certiprof®

BIG DATA

PROFESSIONAL CERTIFICATION

BDPC™ Versión 092022

¿Quién es Certiprof®?

Certiprof® es una entidad certificadora fundada en los Estados Unidos en 2015, ubicada actualmente en Sunrise, Florida.

Nuestra filosofía se basa en la creación de conocimiento en comunidad y para ello su red colaborativa está conformada por:

- **Nuestros Lifelong Learners (LLL)** se identifican como Aprendices Continuos, lo que demuestra su compromiso inquebrantable con el aprendizaje permanente, que es de vital importancia en el mundo digital en constante cambio y expansión de hoy. Independientemente de si ganan o no el examen.
- Las universidades, centros de formación, y facilitadores en todo el mundo forman parte de nuestra red de aliados **ATPs (Authorized TrIAning Partners.)**
- **Los autores (co-creadores)** son expertos de la industria o practicantes que, con su conocimiento, desarrollan contenidos para la creación de nuevas certificaciones que respondan a las necesidades de la industria.
- **Personal Interno:** Nuestro equipo distribuido con operaciones en India, Brasil, Colombia y Estados Unidos está a cargo de superar obstáculos, encontrar soluciones y entregar resultados excepcionales.

Nuestras Afiliaciones

Memberships

itcc
IT CERTIFICATION COUNCIL

Agile Alliance®

ANSI MEMBER

Digital badges issued by

Credly
by Pearson

IT Certification Council - ITCC

Certiprof® es un miembro activo de ITCC.

Una de las ventajas de hacer parte del ITCC es como líderes del sector colaboran entre sí en un formato abierto para explorar maneras nuevas o diferentes formas de hacer negocios que inspiran y fomentan la innovación, estableciendo y compartiendo buenas prácticas que nos permiten extender ese conocimiento a nuestra comunidad.

Certiprof ha contribuido a la elaboración de documentos blancos en el Career Path Ways Taskforce, un grupo de trabajo que se implementó internamente para ofrecer a los estudiantes la oportunidad de saber qué camino tomar después de una certificación.

Algunos de los miembros del ITCC

- **IBM**
- **CISCO**
- **ADOBE**
- **AWS**
- **SAP**
- **GOOGLE**
- **ISACA**

Agile Alliance

Certiprof® es un miembro corporativo de Agile Alliance.

Al unirnos al programa corporativo Agile Alliance, continuamos empoderando a las personas ayudándolas a alcanzar su potencial a través de la educación. Cada día, brindamos más herramientas y recursos que permiten a nuestros socios formar profesionales que buscan mejorar su desarrollo profesional y sus habilidades.

https://www.agilealliance.org/organizations/certiprof/

Credly

Esta alianza permite que las personas y empresas certificadas con Certiprof® cuenten con una distinción a nivel mundial a través de un distintivo digital.

Credly es el emisor de insignias más importante del mundo y empresas líderes en tecnología como IBM, Microsoft, PMI,Nokia, la Universidad de Stanford, entre otras, emiten sus insignias con Credly.

Empresas que emiten insignias de validación de conocimiento con Credly:

- **IBM**
- **Microsoft**
- **PMI**
- **Universidad de Stanford**
- **Certiprof**

Insignias Digitales

Insignias Digitales:
¿Qué Son?

Según el estudio del IT Certification Council (ITCC), años atrás, la gente sabía muy poco sobre las insignias digitales. Hoy, grandes empresas e instituciones educativas de todo el mundo expiden insignias.

Las insignias digitales contienen metadatos detallados sobre quién las ha obtenido, las competencias requeridas y la organización que las ha expedido. Algunas insignias incluso están vinculadas a las actividades necesarias para obtenerlas.

Para las empresas e instituciones educativas, las insignias y la información que proporcionan son tan importantes que muchas decisiones, como las de contratación o admisión, se basan en los datos que aportan.

¿Por qué son importantes?

- **Facilidad de Compartir y Verificar Logros:**

  Las insignias digitales permiten a los profesionales mostrar y verificar sus logros de manera instantánea y global. Según un informe de Credly, **los perfiles de LinkedIn con insignias digitales reciben un 40% más de atención por parte de reclutadores y empleadores.**

- **Visibilidad en Plataformas Digitales:**

  En una encuesta realizada por Pearson y Credly, el **85%** de los usuarios que obtuvieron insignias digitales **las compartieron en LinkedIn,** y el **75%** reportó que esto mejoró su **credibilidad profesional en sus redes.** Además, el **76%** de los empleadores encuestados afirmó que las insignias digitales les ayudan a identificar rápidamente habilidades específicas.

¿Por qué son importantes?

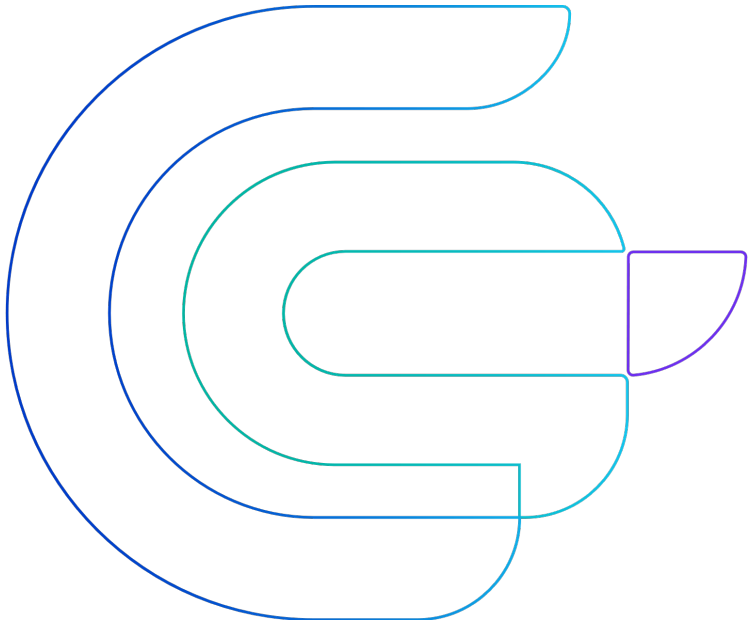
- **Impacto en la Contratación:**

  Un estudio de la **Asociación Internacional de Gestión de Proyectos (PMI)** encontró que los candidatos que muestran insignias digitales de gestión de proyectos tienen **un 60%** más de probabilidades de ser contratados en comparación con aquellos que solo mencionan sus habilidades sin verificación digital.

¿Por qué son importantes?

- **Empoderamiento de la Marca Personal:**

La visibilidad y verificación instantánea proporcionada por las insignias digitales permiten a los profesionales no solo demostrar sus habilidades, sino también construir una marca personal fuerte. Según un estudio de LinkedIn, los profesionales que utilizan insignias digitales tienen un 24% más de probabilidades de avanzar en sus carreras.
La certificación y las insignias digitales no son solo una validación del conocimiento, sino también una herramienta poderosa para la mejora continua y la empleabilidad. En un mundo donde el aprendizaje permanente se ha convertido en la norma, estas credenciales son clave para el desarrollo profesional y la competitividad en el mercado laboral global.



No todas las insignias son iguales, y en **Certiprof,** estamos comprometidos con ofrecerte más que un simple reconocimiento digital. Al obtener una insignia emitida por certiprof, estarás recibiendo una validación de tu conocimiento respaldada por una de las entidades líderes en certificación profesional a nivel mundial.

Da el siguiente paso y obtén la insignia que te abrirá puertas y te posicionará como un experto en tu campo.

¿Por qué es importante obtener su certificado?

- **Prueba de experiencia:** Su certificado es un reconocimiento formal de las habilidades y conocimientos que ha adquirido. Sirve como prueba verificable de sus cualificaciones y demuestra su compromiso con la excelencia en su campo.
- **Credibilidad y reconocimiento:** En el competitivo mercado laboral actual, las empresas y los compañeros valoran las credenciales que le distinguen de los demás. Un certificado de una institución reconocida, como Certiprof, proporciona credibilidad instantánea e impulsa su reputación profesional.
- **Avance profesional:** Tener tu certificado puede abrirte las puertas a nuevas oportunidades. Ya se trate de un ascenso, un aumento de sueldo o un nuevo puesto de trabajo, las certificaciones son un factor diferenciador clave que los empleadores tienen en cuenta a la hora de evaluar a los candidatos.

¿Por qué es importante obtener su certificado?

- **Oportunidades de establecer contactos:** Poseer un certificado le conecta con una red de profesionales certificados. Muchas organizaciones cuentan con grupos de antiguos alumnos o de trabajo en red en los que puede compartir experiencias, intercambiar ideas y ampliar su círculo profesional.

- **Logro personal:** Obtener una certificación es un logro importante, y su certificado es un recordatorio tangible del trabajo duro, la dedicación y el progreso que ha realizado. Es algo de lo que puede sentirse orgulloso y mostrar a los demás.

Insignia

Earn this Badge

Big Data Professional Certificate - BCPC

Issued by Certiprof

Earners of the Big Data Professional Certificate have the skills to analyze data that is complex because of its volume and variability, understanding the importance of analysis of data and how you can develop ideas that lead to better business decisions and strategic movements. They are able to identify problems in an understandable way using Big Data, to provide useful solutions with a large amount of information and with data that can be shaped or tested in any way deemed appropriate

Learn more

Certification | $ Paid

Skills

Analytics And Big Data | Analyze Data | Big Data | Business Analyst

https://www.credly.com/org/certiprof/badge/big-data-professional-certificate-bcpc

Aprendizaje Permanente

- Certiprof ha creado una insignia especial para reconocer a los aprendices constantes.
- Para el 2024, se han emitido más de 1,000,000 de estas insignias en más de 11 idiomas.

Propósito y Filosofía

- Esta insignia está destinada a personas que creen firmemente en que la educación puede cambiar vidas y transformar el mundo.
- La filosofía detrás de la insignia es promover el compromiso con el aprendizaje continuo a lo largo de la vida.

Acceso y Obtención de la Insignia

- La insignia de Lifelong Learning se entrega sin costo a aquellos que se identifican con este enfoque de aprendizaje.
- Cualquier persona que se considere un aprendiz constante puede reclamar su insignia visitando:

https://certiprof.com/pages/certiprof-lifelong-learning

COMPARTE Y VERIFICA TUS LOGROS DE APRENDIZAJE FÁCILMENTE

#BDPC #certiprof

certiprof®

Conceptos Iniciales

Datos Masivos

Los macrodatos son en esencia una inmensa cantidad de datos que pueden tener cierto grado de complejidad y que son masivos por su volumen, lo cual hace que no necesariamente se puedan analizar de una forma tradicional.

Estos datos pueden obtenerse de diversas fuentes y pueden tener tanto carácter cualitativo como cuantitativo, así como variar en cuanto a sus estructuras y por esta razón es que el requerimiento de técnicas especializadas para su análisis es algo que va de la mano para su análisis, identificar patrones y tomar decisiones con base en ellos.

Si bien se podría pensar que el volumen de datos no ha crecido tan exponencialmente desde antes de la década de los 2010, siendo que se están manejando tasas tan inmensas que ya en esta década de los 2020s se habla de Zettabytes (3 escalas arriba de los Terabytes que son lo que podemos encontrar en los discos duros de nuestros equipos con gran almacenamiento, o sea 1000 millones de terabytes), lo cierto es que antes de 2010 ya se vio reflejado que el crecimiento de los datos superó lo previsto por la ley de Moore. Esta ley se postuló en la década de los 1960s donde se conceptualizaba que el procesamiento computacional aumentaría a mayor velocidad, sería más pequeño el hardware requerido para su procesamiento y su eficiencia aumentaría con el pasar de tiempo (particularmente enfocándose en el número de transistores en los microchips)... el asunto es, el crecimiento de los datos supera a la ley de Moore desde la segunda mitad de la década de los 2000 en comparación al escalado de transistores (Kachris & Tomkos, 2015).

Datos Masivos

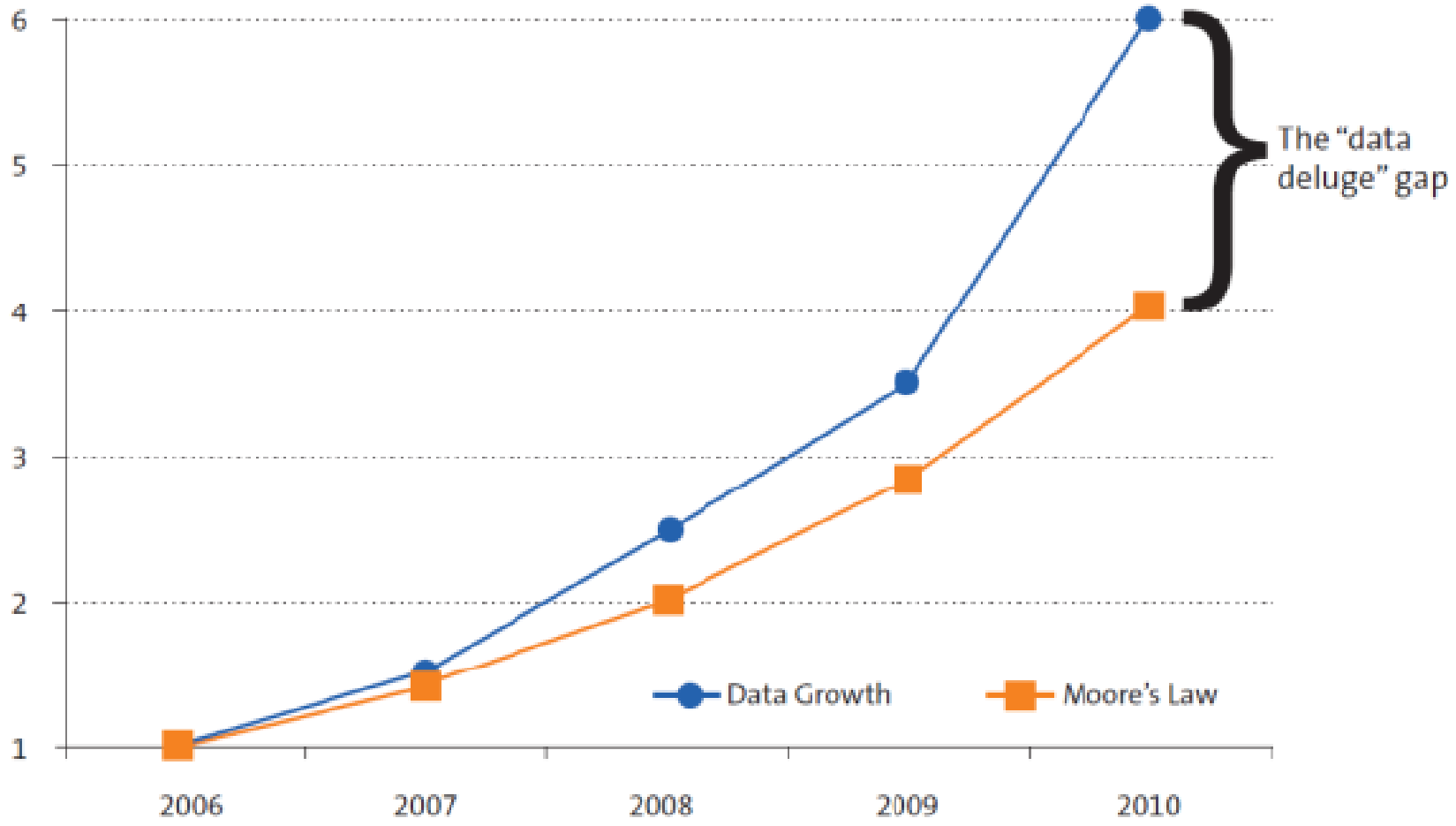
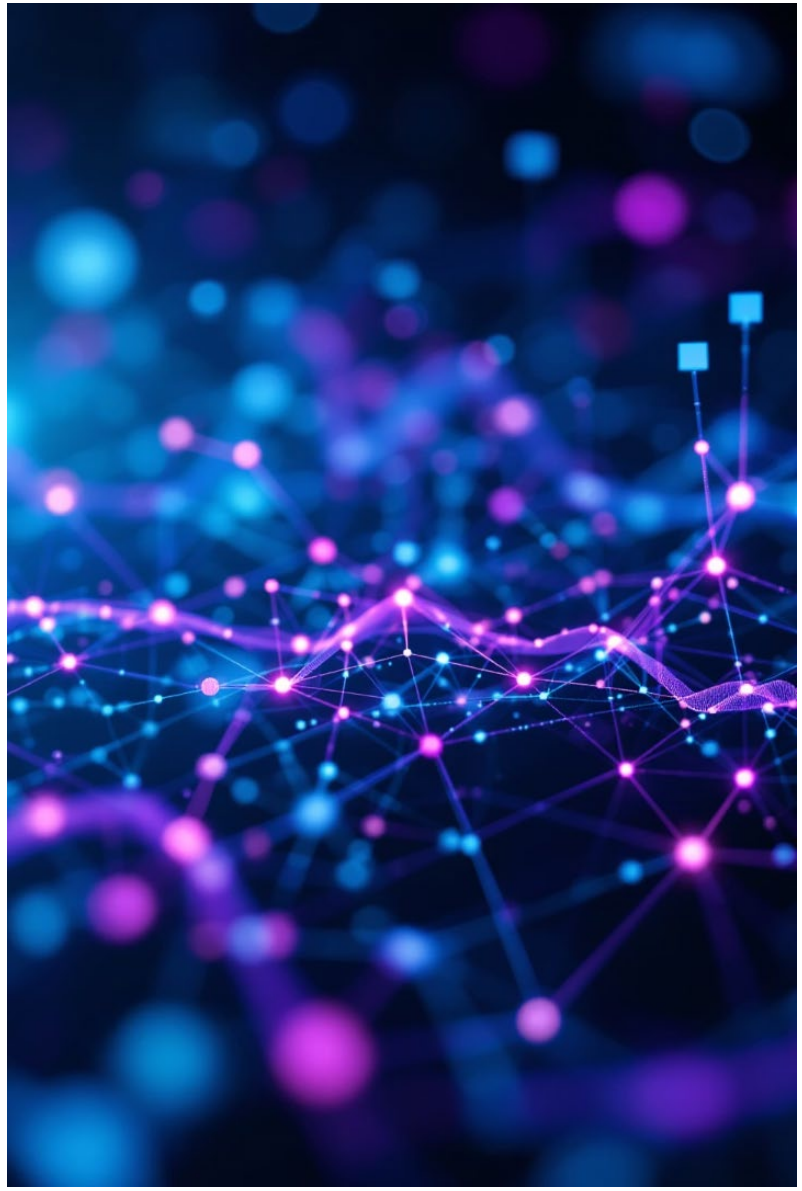


Figura 1. La brecha del "diluvio de datos". Obtenido de (Kachris & Tomkos, 2015).

Big Data

El concepto de Big Data puede partir a raíz de lo que expone Gartner, donde se indica que "son activos de información de gran volumen, alta velocidad y/o gran variedad que exigen formas innovadoras y rentables de procesamiento de información que permiten una mejor comprensión, toma de decisiones y automatización de procesos". (Gartner, s.f.). Para esto, y como se presentaba con anterioridad, es necesaria una confluencia de prácticas, estándares, tecnologías y arquitecturas para la gestión y análisis de volúmenes de datos que no pueden ser tratados de manera convencional, y que superan los límites y capacidades de las herramientas de software usadas para la captura, gestión y procesamiento de datos.



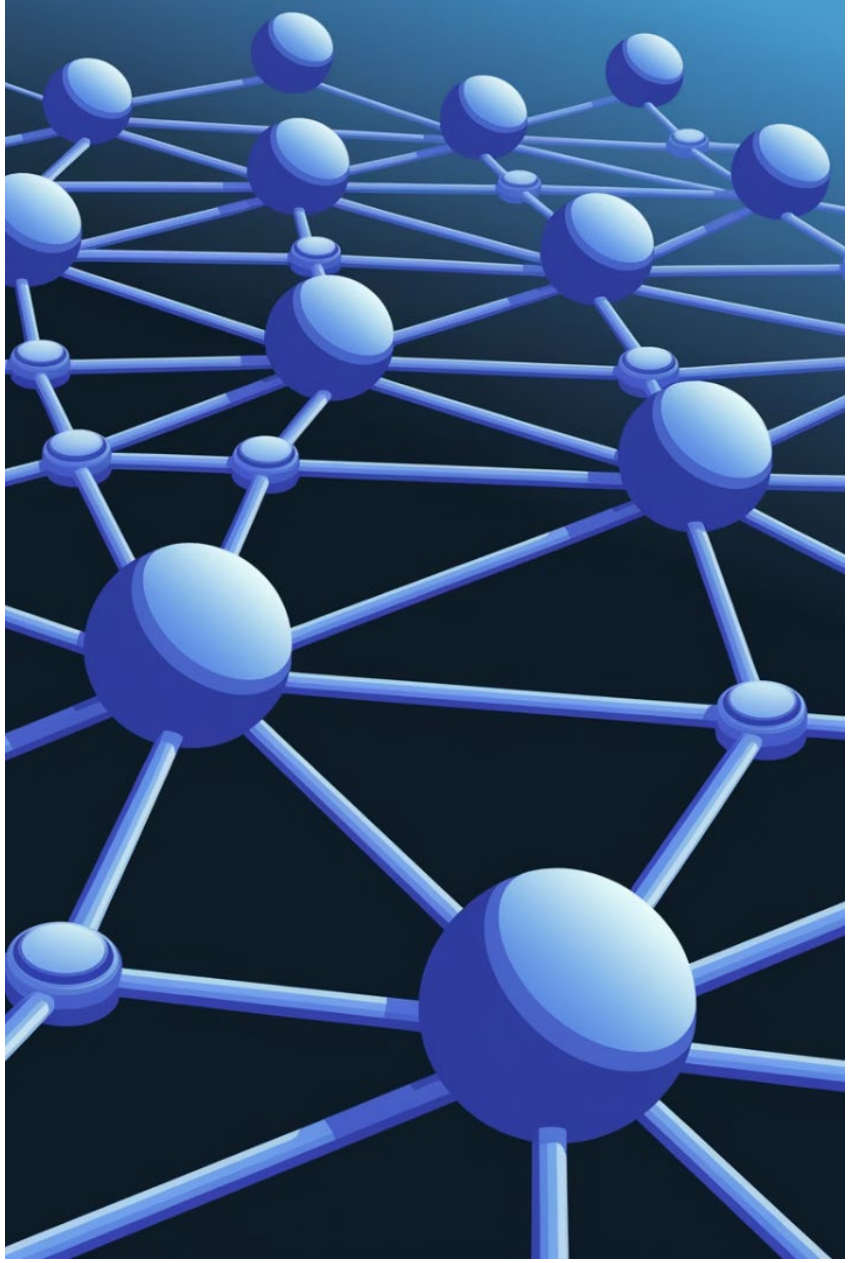
Big Data

Llevado un poco a nuestro diario vivir, es algo que se orienta más a la obtención de insights del comportamiento humano en aras de permitir tomar mejores decisiones y actuar con mayor velocidad basados en datos que son de amplia variedad y de un alto volumen. Para entender las características del Big Data, existen las llamadas 7Vs, pero esas se mencionarán en el siguiente apartado ya que lo que concierne aquí con mayor relevancia es el hecho de que el avance tecnológico ha conllevado a que los datos sobre nosotros y sobre lo que hacemos aumenten como lo puede ser el hecho de que hoy día almacenamos con total normalidad miles y miles de fotos solo en un dispositivo móvil cuando en la década de los 2000 uno debía seleccionar qué fotos permanecían almacenadas. También viene siendo cierto que los dispositivos que se denominan como "inteligentes" ya vienen con capacidad de conectarse entre sí y esto permite que se capten aún más datos, siendo que si tenemos dos o más dispositivos de estas características en nuestros hogares y están conectados entre sí pues lo natural es que estemos alimentándolos con muchos datos que después requerirán de unos esfuerzos importante para su análisis.

Las 7Vs del Big Data

Originalmente, el Big Data se compuso de una definición que partía de 3Vs, las cuales eran el volumen, la velocidad y la variedad (Russom, 2011). No obstante, para el momento en que se planteó apenas se hablaba de terabytes y ya hemos discutido que hoy día hablamos de zettabytes, lo cual no solo es una muestra de que ha habido una evolución a nivel de volumen de datos, sino que seguramente también de velocidad de generación y captación de los mismos, y por tanto de su variedad (se encontrarán siempre muchos datos que tendrán tipos diferentes entre estructurados, semiestructurados, no estructurados o todos los anteriores).

Posteriormente, siendo notorio que hubo avances en el Big Data y su caracterización, se unieron el valor y la veracidad como las nuevas Vs del Big Data para así formar 5Vs (Ashaari et al., 2021). De estas primero se incluye el valor respecto a la extracción del mismo a partir de los datos captados para que obtenga algún tipo de beneficio a raíz de esto, y luego se incluye la veracidad para que así se pueda tener en consideración la importancia de la calidad de los datos captados sobre la mesa para así generar un aumento en la confianza de la data utilizada para la toma de decisiones (Fosso Wamba et al., 2015).



Las 7Vs del Big Data

Para las últimas 2Vs del Big Data de las cuales se hará presentación existe una particularidad y es el hecho de que si bien está consensuado el hecho de que una de las dos pueda mantenerse como visualización (Chaudhari, 2019), la otra sí varíe dependiendo el autor y es que la visualización de los datos se refiere a visualizar la relación que existe entre la data para poder dar una representación gráfica a unas series de data que pueden ser complejas entender por separado, para así comunicar de mejor manera un mensaje que provea todas las Vs previamente discutidas.

A pesar delo anterior, el motivo por el cual la 7ma V del Big Data varía dependiendo los autores es porque depende bajo qué espectro u óptica lo estén considerando para determinar su peso dentro de las 7Vs.

En el caso de Chaudhari et al. (2019) & Ashaari et al. (2021) es la variabilidad que se diferencia de la variedad en cuanto a que "si la importancia de la información se transforma continuamente, puede afectar la homogeneización de su información", mientras que para Khan et al. (2014) la volatilidad cobra valor porque es algo que emerge gracias a las 3 primeras Vs del Big Data y puede llegar a afectar factores como los periodos de retención de datos, así como el gasto de almacenamiento y seguridad de los mismos (aunque también incluyen la viabilidad en vez de visualización, la cual se asemeja a la exactitud de los datos frente a su uso previsto, o sea, su uso eficaz).

Las 7Vs del Big Data

Como se puede apreciar, la discusión sobre las 7Vs de Big data tiene un común denominador frente a las primeras 5Vs (volumen, velocidad, variedad, valor y veracidad), pero pueden haber alteraciones frente a las últimas 2Vs, de las cuales resaltan la visualización y la 7ma quedaría a juicio del uso de implementación de un proyecto de Big Data que se desee llevar a cabo puesto que la volatilidad, viabilidad y variabilidad no son más importantes una que la otra, sino más bien complementarias y aspectos a considerar a la hora de tratar con los datos de esta magnitud.

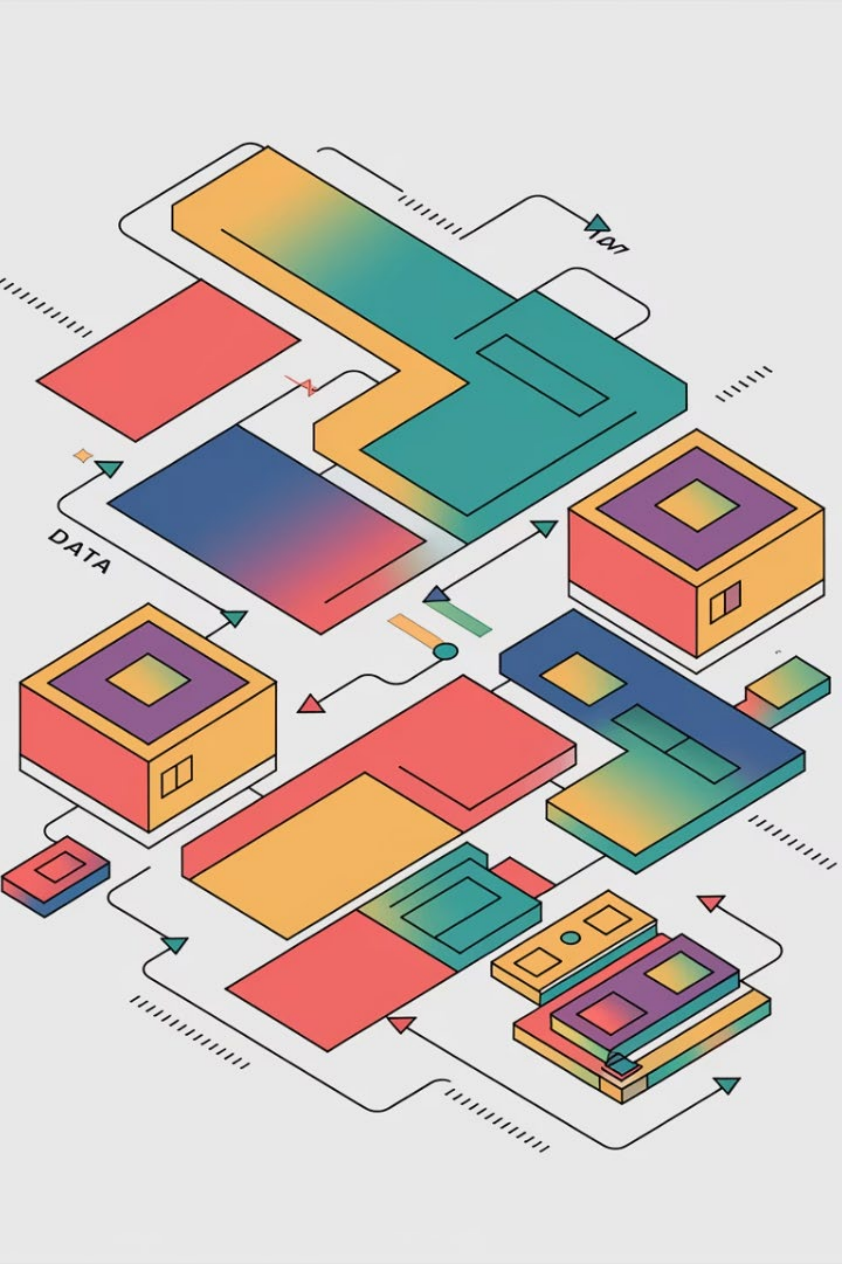
Las 7Vs del Big Data

Figura 2. Las 7Vs del Big Data.

MapReduce

Map reduce es una técnica por la cual es posible dividir los datos en porciones pequeñas para luego procesarlas por separado, como paradigma de programación se ha implementado desde herramientas Open Source como Hadoop.

Hoy día se implementa sobre otros framewoks de datos. Resulta ser que en Big Data convencional que conocemos no existe un mecanismo como tal que permita la identificación de la data sensible tanto estructurada como no estructurada y esto hace que sea necesario contar con otros tipos de mecanismos que aporten a este tipo de protección individual de los datos, por lo que se dio un trabajo orientado a incluir una capa de seguridad al modelo de MapReduce tradicional enfocándose en la individualidad de los datos y el output obtenido del modelo para así proveer la seguridad necesaria con ayuda de criptografía (Gudditti & Venkata, 2021).



MapReduce

Partiendo del hecho de que MapReduce es una plataforma de programación embebida en varios framewoks, como lo son el caso de Hadoop, Spark o Python, para así poder llegar a alcanzar ese análisis masivo de data de una forma paralela (Li et al., 2020), se incluye que aparte de contar entre las tecnologías consideradas para el procesamiento de Big Data como lo son el MapReduce y Hadoop, también se habla de MongoDB y Cassandra como tecnologías comunes para poder atender desafíos comunes como lo son el captar, organizar, almacenar, buscar, compartir, transferir y analizar el Big Data, e inclusive se han propuesto modelos de alta complejidad como el denominado MR-MVPP o "Construcción basada en MapReduce del Plan de procesamiento de vistas múltiples" por sus siglas en inglés (Azgomi & Sohrabi, 2021).

Dentro de los modelos donde se integra el MapReduce puede existir una gran variedad y esto aumenta su complejidad de implementación, pero enriquece la salida de datos final. Sin embargo, una simple representación gráfica de lo que es un MapReduce se presenta en la infografía al respecto donde hay una entrada y una salida de datos, los mapas representan entradas pares o un conjunto de inputs pares en cuanto a clave/valor y luego se dan las tareas de reducción para así lograr el output final de datos.

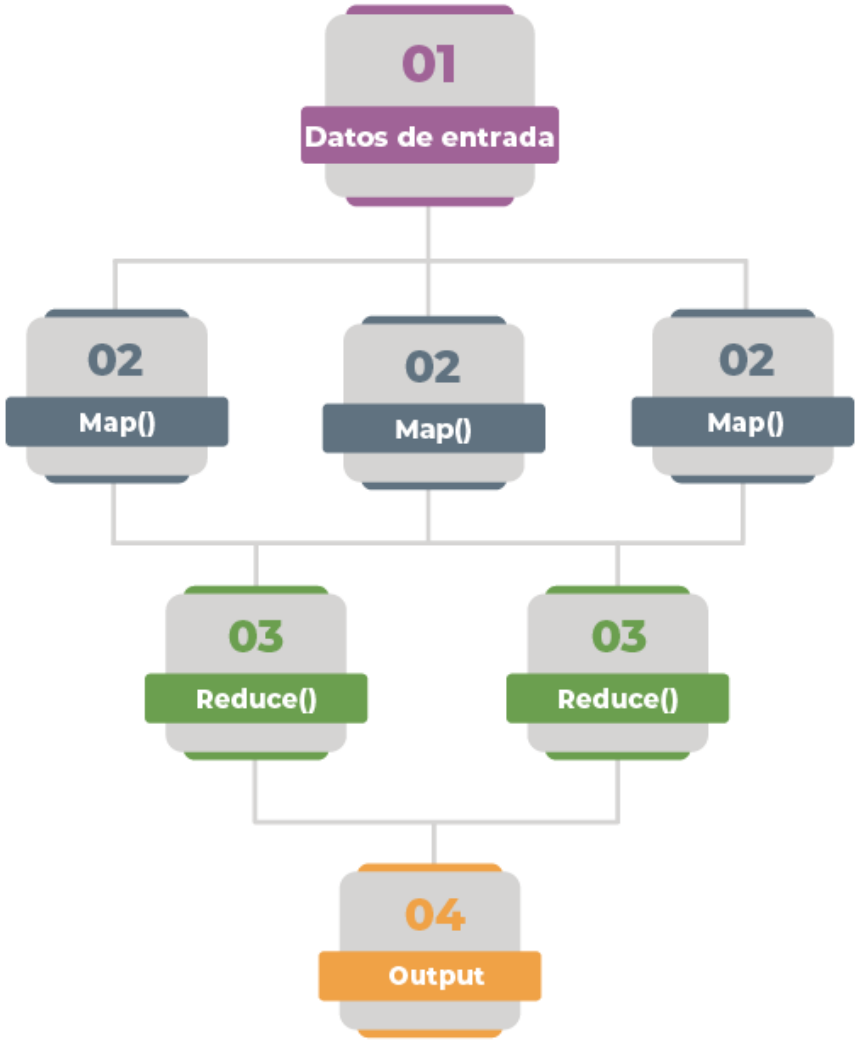


Figura 3. Proceso tradicional base del MapReduce.

Datos, Información y Conocimiento

Los Datos

Definición: Los datos se pueden definir bajo lo que se postula en el diccionario de Cambridge, donde se les refiere como "información, especialmente hechos o números, recopilada para ser examinada, considerada y utilizada para ayudar a tomar decisiones", o también como "información en una forma electrónica que puede ser almacenada y procesada por un computador". En cualquiera de los dos casos, es una definición muy clara de los datos y de los requerimientos necesarios para así poder procesarlos y tomar una decisión basada en ellos.

Semánticas del valor representado (Número, posición, código, imagen): De acuerdo con Gartner (s.f.) el modelo semántico de los datos es "un método de organización de datos que refleja el significado básico de los elementos de datos y las relaciones entre ellos. Esta organización facilita el desarrollo de programas de aplicación y el mantenimiento de la coherencia de los datos cuando se actualizan".

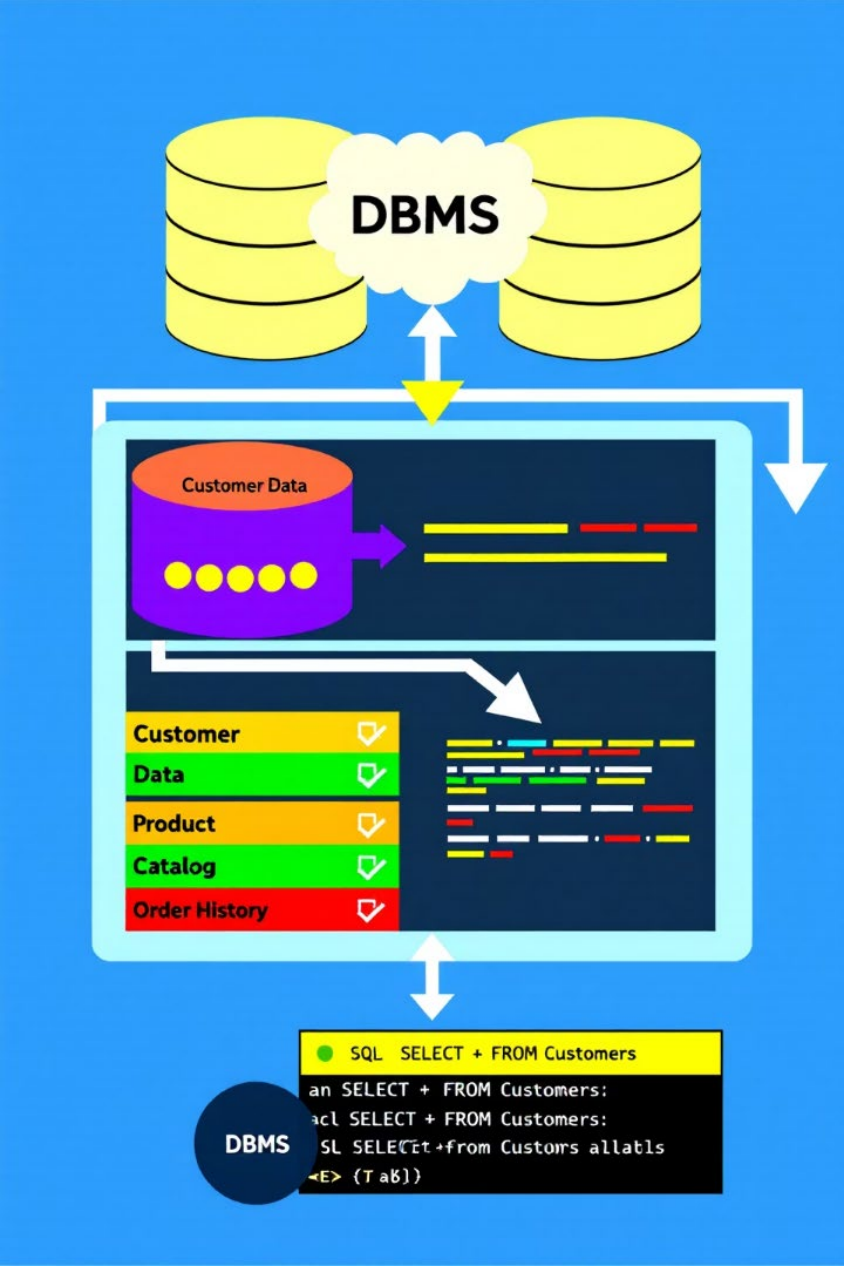
Los Datos

- **Orígenes o fuentes:** Una fuente u origen de dato es un lugar de donde estos son obtenidos, y esta fuente puede ser una base de datos; un archivo plano; un fichero XML; o cualquier tipo de formato que pueda ser legible por un sistema que esté trayendo estos datos. Las fuentes de datos pueden ser variadas y a nivel empresarial un negocio puede manejar perfectamente múltiples fuentes de datos según su objeto de negocio, siendo que habrá una captación de fuentes de datos internas, externas o de acceso abierto. De acuerdo con Oracle (2022) las fuentes de datos pueden apuntar a una base de datos que se encuentre en una ubicación o una máquina específicas en la compañía que esté orientada al procesamiento de la lógica de estos datos.

- **Granularidad:** Dentro del modelo de semántica de datos se emplean 3 tipos de abstracción de acuerdo con la empresa GoodData (2022), los cuales son la clasificación, la agregación y la generalización:
  - **Clasificación:** Esto clasifica diferentes objetos en la realidad objetiva mediante el uso de relaciones de "instancia de", como la creación de grupos de objetos por características similares.
  - **Agregación:** La agregación define un nuevo objeto a partir de un conjunto de objetos que se convierten en sus componentes usando relaciones "tiene un".
  - **Generalización:** La generalización define la relación de un subconjunto entre ocurrencias de dos o más objetos mediante el uso de relaciones "es un".
  - **Ejemplo:** La clasificación puede ser un grupo de estudiantes, la agregación es que tengan nombre, edad, etc., y la generalización es que un profesor sea una generalización de docentes.

Los Datos

- **Calidad:** Valoración conjunta de atributos de origen, confiabilidad y usabilidad del dato.
- **Tipos de datos:**
  - **Estructurados (SQL):** Son datos que se organizan fácilmente en series de filas y columnas y mapeados en campos fijos y predefinidos para su entendimiento, siguiendo un modelo de relaciones por alguien que gestione las bases de datos.
  - **Semiestructurados (JSON, XML, CSV):** Esencialmente es una mezcla entre las categorías de datos estructurados y no estructurados donde se requiere que exista una metadata vista como una lógica simple que permita que su organización sea más sencilla a la hora de su análisis y procesamiento.
  - **No estructurados (Mapas de bits, audio, video):** Son datos que no se pueden contener en bases de datos similares a las de los estructurados ya que la carencia de esas estructuras fijas hace que se requieran unos esfuerzos importantes de preprocesamiento de data, así como el hecho de que suelen ser contenidos comúnmente en lagos de datos (data lakes).



Las Bases de Datos

Las bases de datos son recopilaciones de información o de datos estructurados, manteniendo siempre una forma organizada de su representación para que puedan encontrarse almacenadas electrónicamente. Comúnmente, se les controla con lo que se conoce como DBMS (o sistemas de gestión de bases de datos en español, o también llamados Sistemas Gestores de Bases de datos). Los motores de bases de datos suelen enfocarse en utilizar lenguajes de consulta estructurada (SQL), pero las bases de datos han tenido una evolución importante y se han diversificado sus tipos de acuerdo con como se expone por Oracle (s.f.) y se presenta a continuación:

Las Bases de Datos

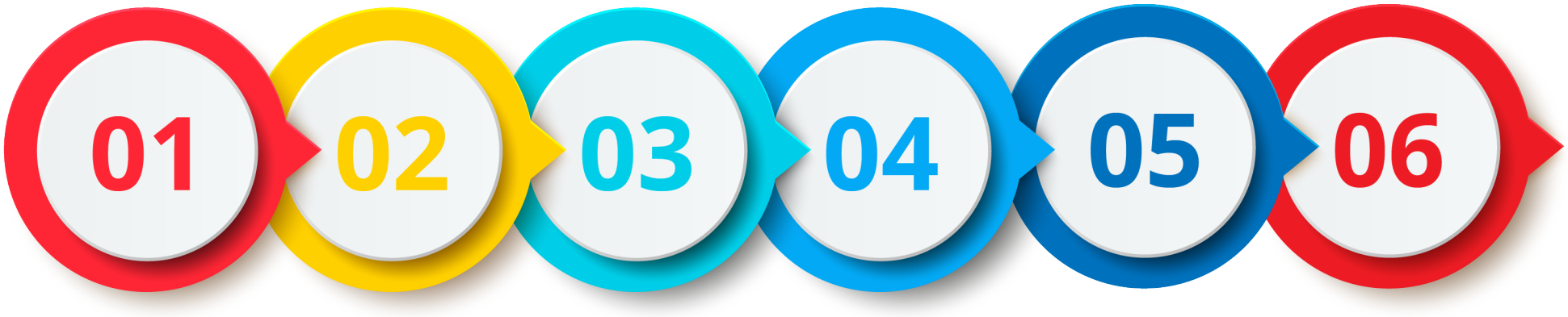


Figura 4. Tipos de bases de datos. Información adaptada de (Oracle, s.f.).

Las Bases de Datos

Bases de datos de código abierto

Aquel cuyo código fuente es de código abierto; tales bases de datos pueden ser bases de datos SQL o NoSQL

Bases de datos en la nube

Es una recopilación de datos, estructurados o no estructurados, que reside en una plataforma de cloud computing privada, pública o híbrida

Base de datos multimodelo

Combinan distintos tipos de modelos de bases de datos en un único servidor integrado, lo que les permite incorporar distintos tipos de datos

Bases de datos de documentos/JSON

Diseñadas para almacenar, recuperar y gestionar información orientada a los documentos, las bases de datos de documentos son una forma moderna de almacenar los datos en formato JSON en lugar de hacerlo en filas y columnas

Bases de datos de autogestión (o autónomas)

Están basadas en la nube y utilizan machine learning para automatizar el ajuste de la base de datos, la seguridad, las copias de seguridad, las actualizaciones y otras tareas de gestión rutinarias que tradicionalmente realizan los administradores de bases de datos

Figura 4. Tipos de bases de datos. Información adaptada de (Oracle, s.f.).

Modelos de Datos

Patrones de estructuración de datos estandarizados como definición de una base de datos de acuerdo con las descripciones formales en su sistema de información y según los requisitos del sistema de gestión de base de datos que se aplicará. Los modelos de datos son el modelado de la descripción de los datos, la semántica de los datos y las restricciones de consistencia de los datos, de donde podemos extraer 4 tipos de modelos de datos, los cuales son: Modelo de datos relacional, modelo de datos semiestructurado, modelo de datos de entidad-relación, y modelo de datos basado en objetos (JavaTPoint, s.f.).

En el modelado de datos tradicionalmente se pueden distinguir 3 niveles o pasos para su ejecución, los cuales son: 1) Modelado conceptual de datos donde se identifica y describe las entidades y su relación, así como la notación de sus descubrimientos; 2) Modelado lógico de datos donde se definen las tablas de la base de datos de acuerdo con la relación subyacente del modelo que se haya planteado a nivel conceptual; y 3) Modelado físico de datos donde se da la optimización de la base de datos para obtener el máximo rendimiento posible (Flanders & Jannidis, 2015).

Modelos de Datos

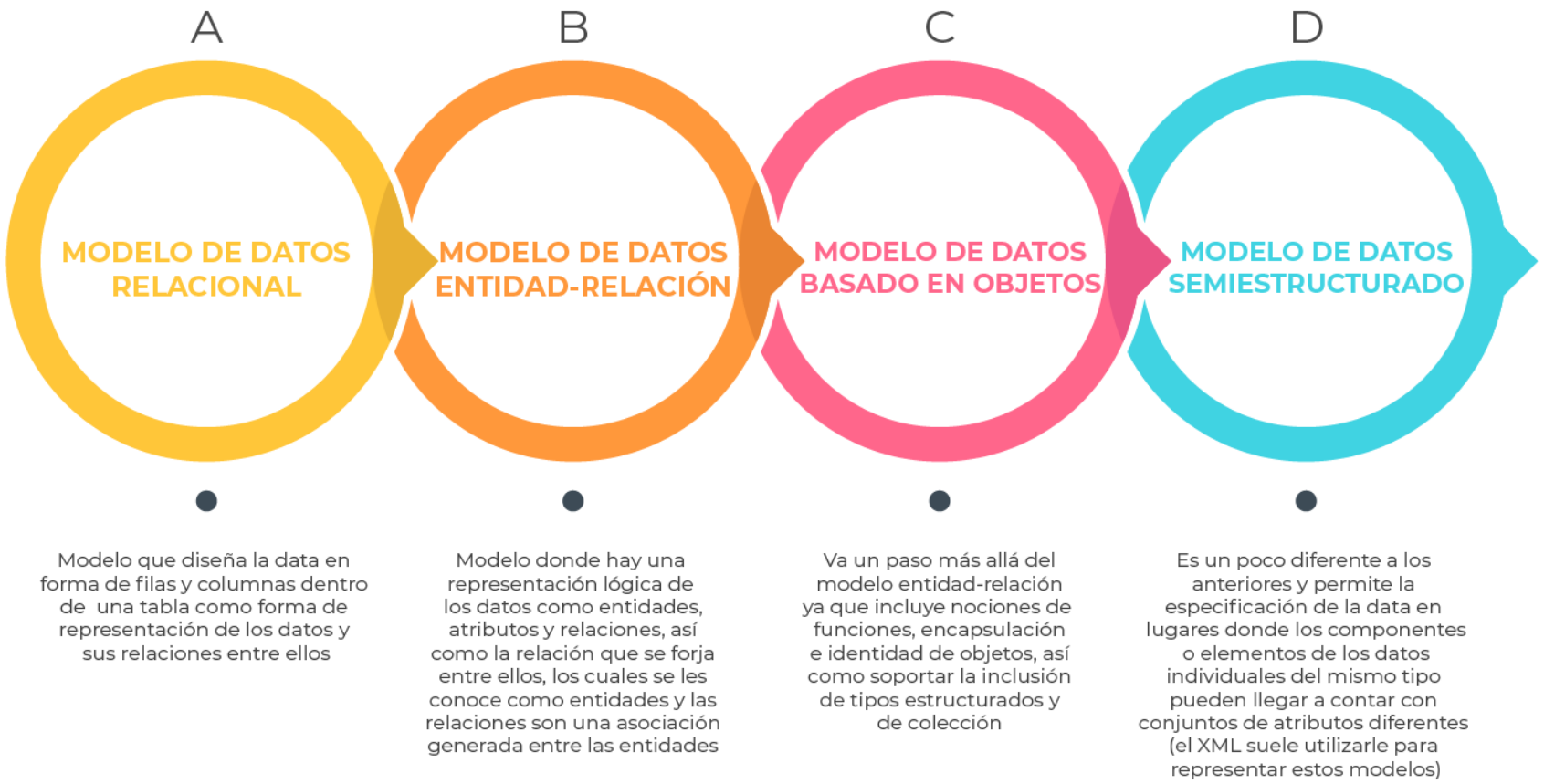


Figura 5. Tipos de modelos de datos. Información adaptada de (JavaTPoint, s.f.)

Procesamiento de Datos

Realización sistemática de operaciones de transformación sobre datos. Este procesamiento de datos requiere que se tengan en cuenta las diversas Vs del Big Data que ya se han discutido, así como el hecho de que aparte de saber que puede variar el volumen de datos a procesar, la complejidad de la operación del procesamiento de datos es algo crítico ya que debe velarse por el hecho de que la capacidad y la tecnología computacional esté con la posibilidad de realizar dicho procesamiento de datos. Asimismo, al ser un proceso que tiene por objeto servir a un objeto organizacional, se tiene que considerar cuáles son las habilidades técnicas requeridas para esta operación, el personal necesario y las restricciones de tiempo para la obtención del resultado final. Este procesamiento de datos cuenta con unas fases como se presenta en la siguiente infografía.

Procesamiento de Datos

Figura 6. Fases o etapas del procesamiento de datos. Información adaptada de (Cruz et al., s.f.; JavaTPoint, s.f.).

Procesamiento de Datos

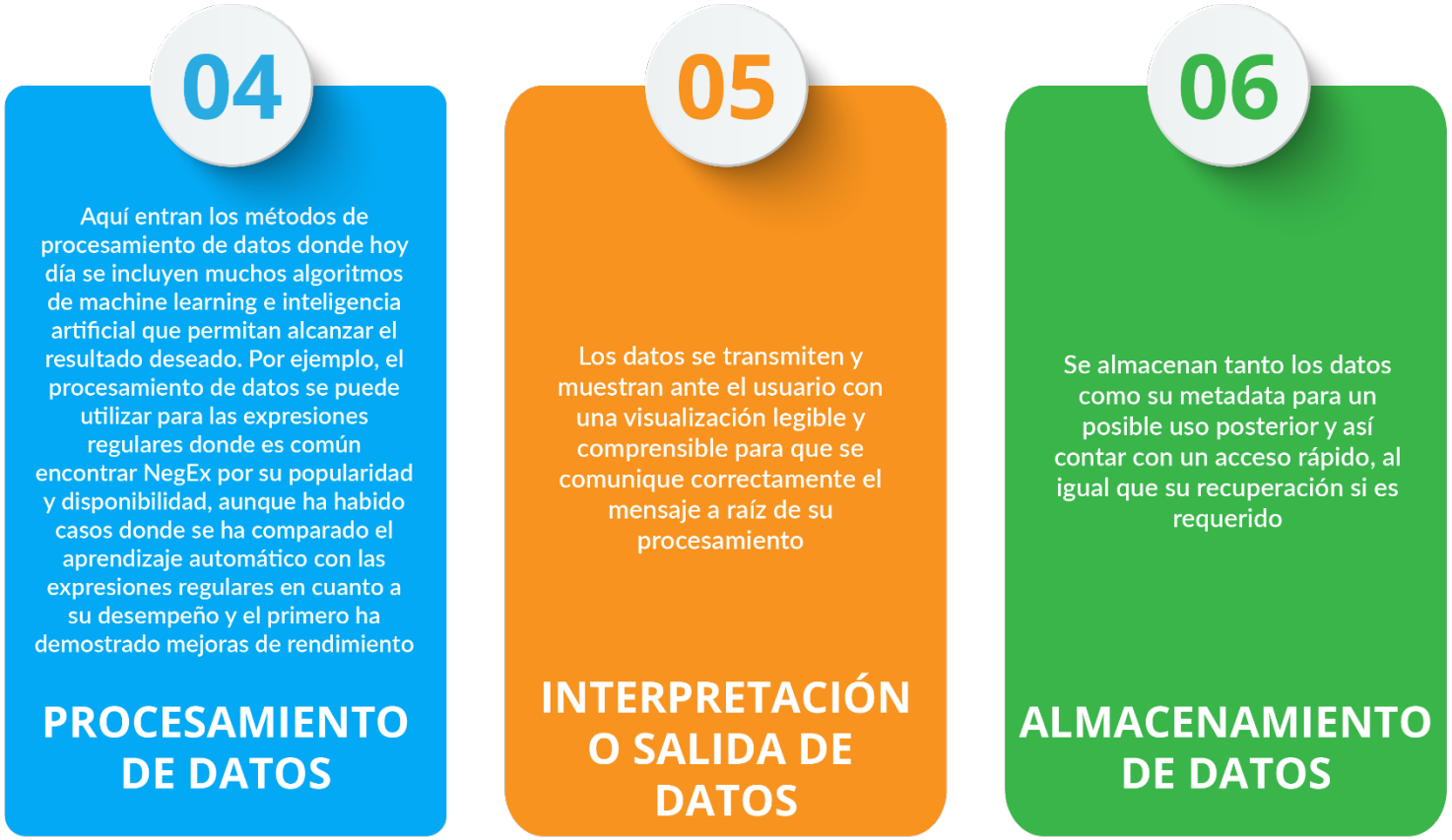
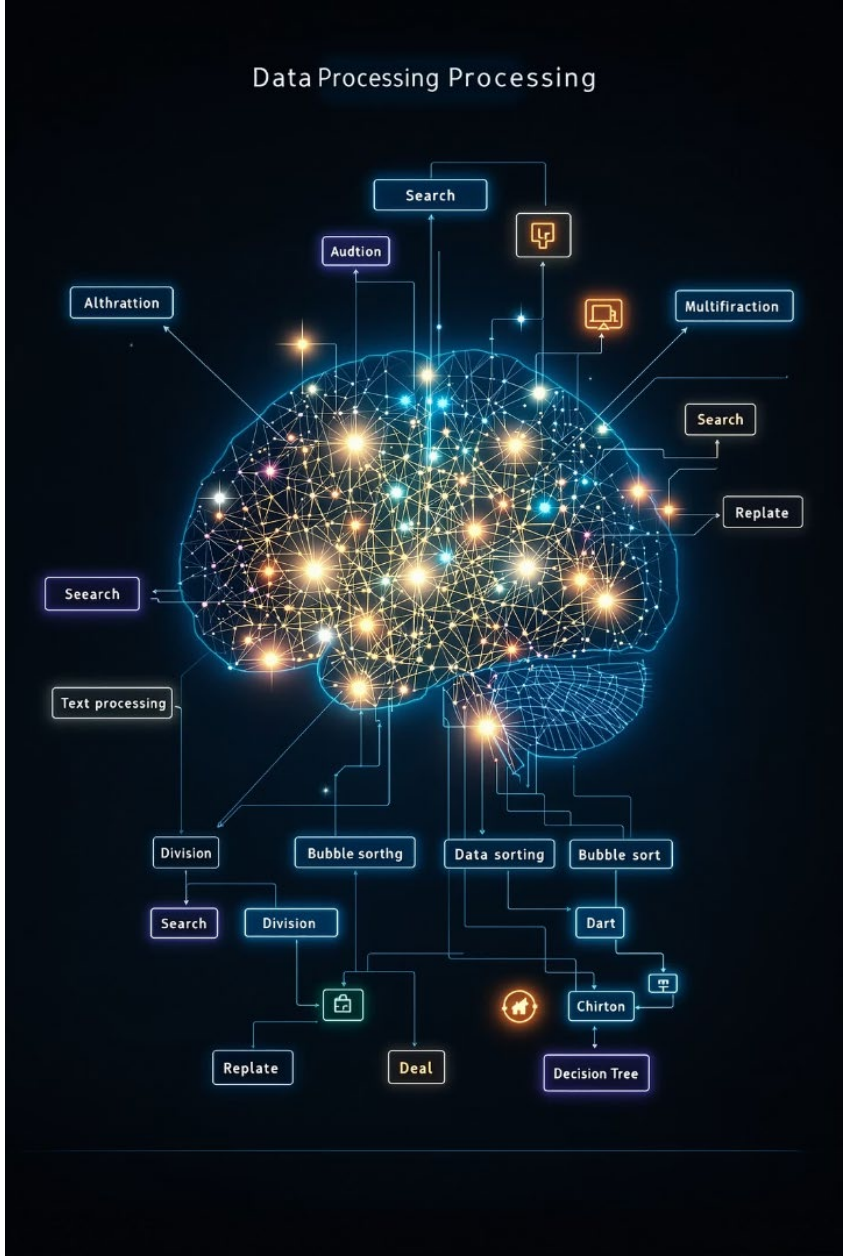


Figura 6. Fases o etapas del procesamiento de datos. Información adaptada de (Cruz et al., s.f.; JavaTPoint, s.f.).



Procesamiento de Datos

Ejemplos de procesamiento de datos: Operaciones aritméticas o lógicas sobre datos, fusión o clasificación de datos u operaciones sobre texto, tales como la como edición, clasificación, fusión, almacenamiento, recuperación, visualización o impresión.

Nota: El término procesamiento de datos no debe usarse como sinónimo de procesamiento de información.

Información

BDPC™ Versión 092022

Información

Es el valor de referencia o propiedad agregada de los datos que tiene sentido como como salida (o output) de un proceso de datos. Resulta de las operaciones que se realizan sobre los datos, perdiendo generalidad de detalle. La información entonces comprende lo que es la presentación de data procesada y organizada en un contexto que sea significativo para algo determinado, sabiendo que la data agrupada conlleva colectivamente consigo un significado lógico que aporta a la toma de decisiones.

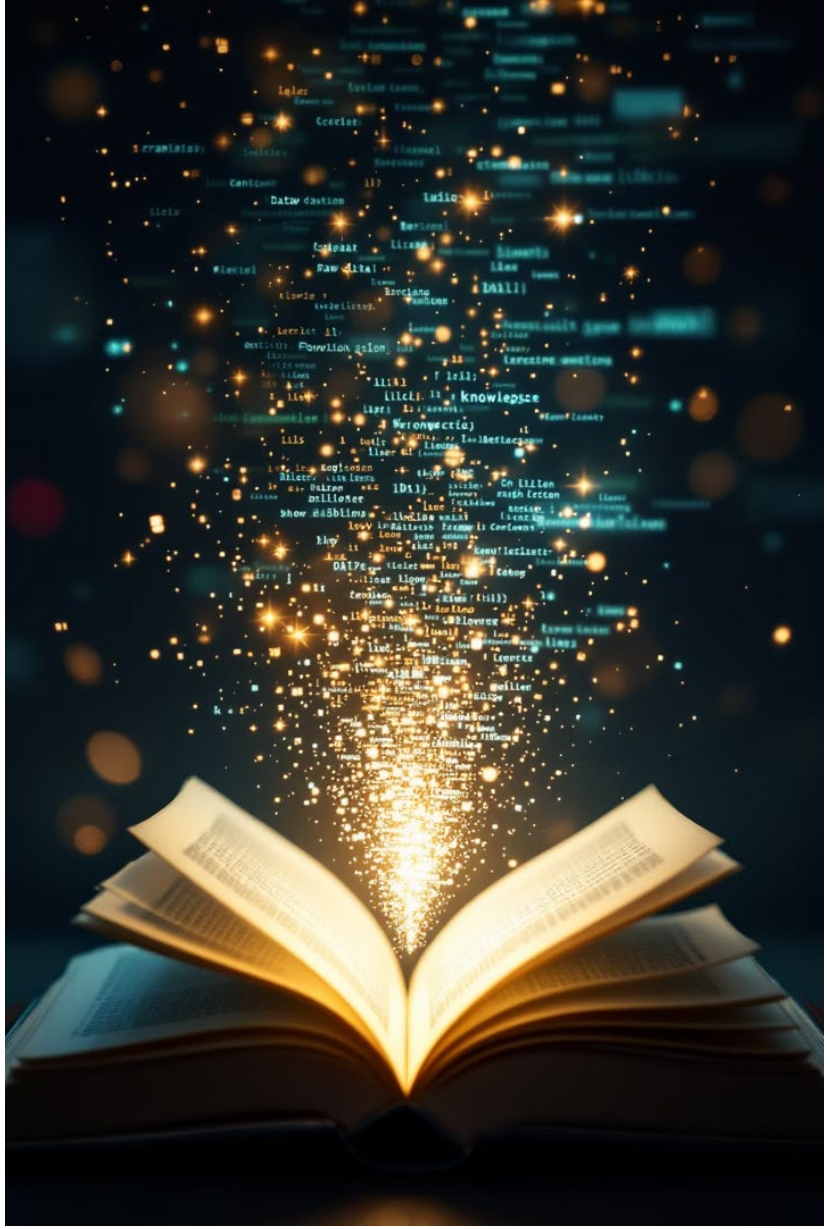
Características Dependientes del Proceso Utilizado, el Dominio y la Interpretación:

Figura 7. Características dependientes del proceso utilizado, el dominio y la interpretación.

Clústeres de Datos o Sistemas Distribuidos

Figura 8. Características de los clústeres de datos o sistemas distribuidos.

Conocimiento



Conocimiento

Abstracción conceptual que regula el proceso que se deriva de la transformación de los datos en información. Es totalmente indivisible de la lógica de contexto (o de negocio) y objetivo a alto nivel, para lo cual se asume como un grado de generalización conceptual donde existe un equilibrio entre su aplicabilidad y su entendimiento. El conocimiento entonces es una fase posterior a las que hemos visto con anterioridad, que son data e información, y para verlo mejor representado se maneja la siguiente infografía donde se expresan los pasos del modelo DIKW (Data-Información-Conocimiento-Sabiduría en español)

Conocimiento

Abstracción conceptual que regula el proceso que se deriva de la transformación de los datos en información. Es totalmente indivisible de la lógica de contexto (o de negocio) y objetivo a alto nivel, para lo cual se asume como un grado de generalización conceptual donde existe un equilibrio entre su aplicabilidad y su entendimiento. El conocimiento entonces es una fase posterior a las que hemos visto con anterioridad, que son data e información, y para verlo mejor representado se maneja la siguiente infografía donde se expresan los pasos del modelo DIKW (Data-Información-Conocimiento-Sabiduría en español)

Conocimiento

Figura 9. Modelo DIWK. Información adaptada de (Rao, 2018).

Conocimiento

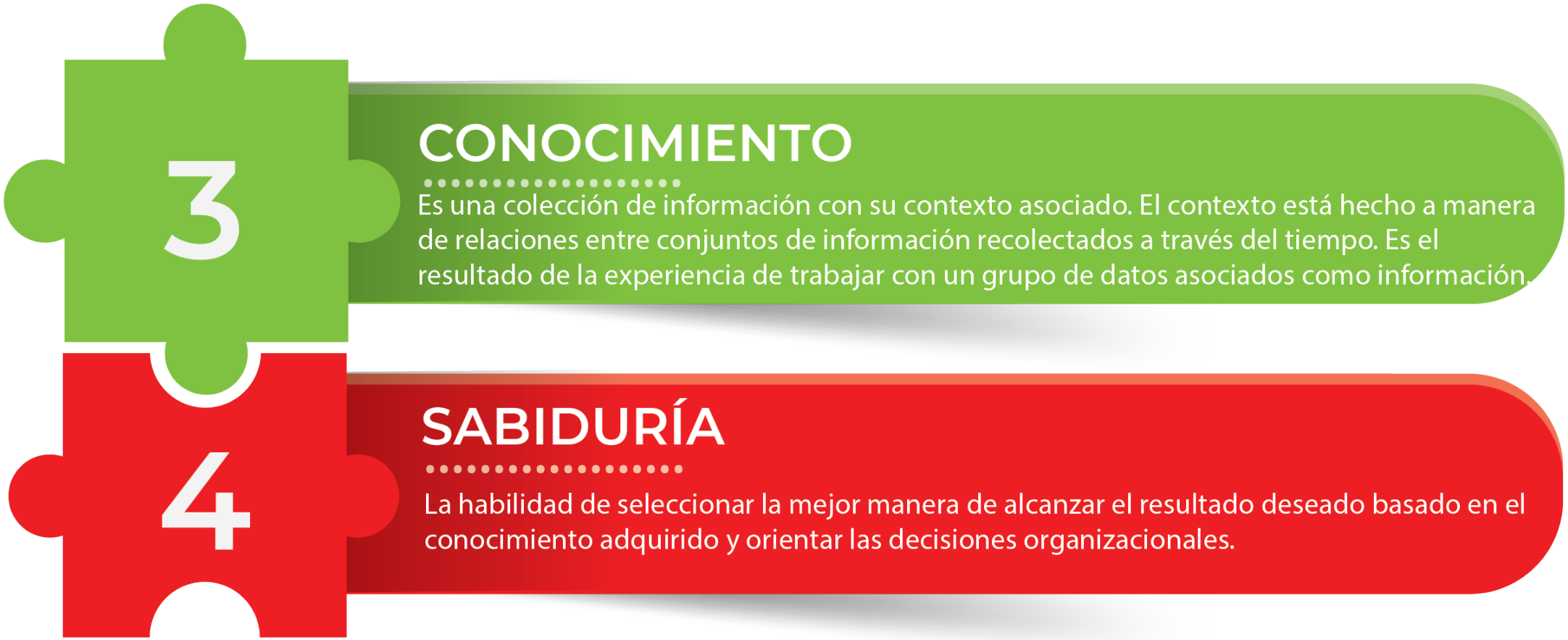


Figura 9. Modelo DIWK. Información adaptada de (Rao, 2018).

Perspectiva, Eficacia y Propósito

A

PERSPECTIVA

Todo conocimiento posee una estructura de interrelaciones con otros conocimientos, bases teóricas y experiencias, lo que representa un re) corte de la realidad interpretada

B

EFICACIA

La información condensada en cono) cimiento posee expectativas sobre sus valores y puntos de apalancamiento sobre los que se debería actuar para modificar la información

C

PROPÓSITO

Existe un objetivo asociado al cono) cimiento que regula la manera de transformar los datos en información

Figura 10. Perspectiva, eficacia y propósito.

La Inteligencia Artificial para Generar Conocimiento

Machine learning o aprendizaje automático es uno de los campos de la informática (o ciencias de la computación) utilizados a la hora de buscar y aprender sobre patrones de la data mediante operaciones matemáticas, supuestos estadísticos y reglas que permitan alcanzar dicho resultado para que en últimas se pueda contar con modelos de carácter predictivo frente a los conjuntos de datos analizados (Singh & Kaushik, 2022).

La Inteligencia Artificial para Generar Conocimiento

Aprendizaje supervisado

El aprendizaje supervisado es una aproximación orientada hacia conjuntos de datos que están etiquetados y que se diseñan para entrenar o supervisar algoritmos que estén enfocados en la clasificación o predicción exacta de resultados (Delua, 2021). Este tipo de aprendizaje se separa en dos tipos, los cuales son la clasificación y la regresión, donde el primero es un algoritmo "para asignar con precisión datos de prueba en categorías específicas" (Delua, 2021), mientras que el segundo busca entender la relación entre variables independientes y dependientes, siendo útiles para la predicción de valores numéricos basados en diferentes grupos de datos.

La Inteligencia Artificial para Generar Conocimiento

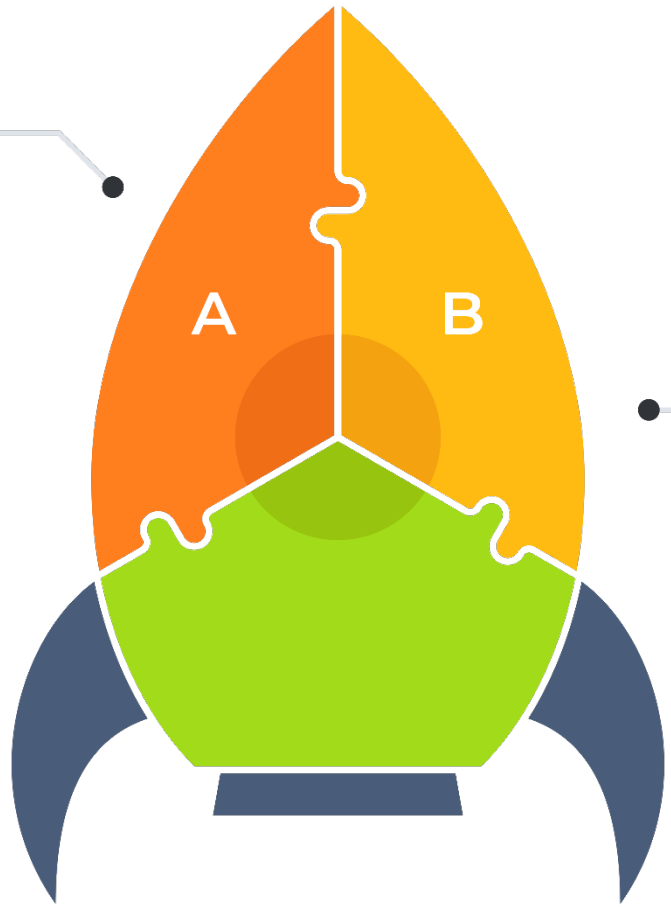
Aprendizaje no supervisado

Es el que utiliza algoritmos para analizar y agrupar grupos de datos no etiquetados para encontrar patrones ocultos sin la necesidad de una intervención humana. Principalmente, se enfocan en 3 tareas que son el clustering, la asociación y la reducción de dimensionalidad. El primero es una técnica para agrupar data no etiquetada de acuerdo con sus similitudes o diferencias, el segundo busca encontrar relaciones entre variables de un conjunto de datos dado, y el tercero se da cuando el número de características (o dimensiones) dentro de un conjunto de datos dado es demasiado alto, así que se busca reducir los inputs de datos a algo que sea manejable (Delua, 2021).

Infografía

APRENDIZAJE SUPERVISADO

Clasificadores lineales, máquinas de vectores de soporte (SVMs), árboles de decisión, redes bayesianas, bosques aleatorios, regresión lineal, regresión logística y re-gresión polinomial



APRENDIZAJE NO SUPERVISADO

Algoritmos de clustering de K-means, deep learning

Figura 11. Algoritmos de aprendizaje supervisado aprendizaje no supervisado.

Evolución de Big Data

Las Tres Grandes Fases de la Evolución de Big Data

Periodo: 1970-2000

FASE 1 BIG DATA

Contenido estructurado basado en SGBD:

1. RDBMS & data warehousing
2. Carga de transferencia de carga y transferencia de datos
3. Procesamiento de datos en línea
4. Cuadros de mando y scorecard
5. Minería de datos y análisis estadístico

Periodo: 2000-2010

FASE 2 BIG DATA

Contenidos no estructurados basados en la web:

1. Recuperación y extracción de información
2. Minería de opinión
3. Análisis de la web y patrones de datos
4. Análisis de redes sociales
5. Análisis espaciotemporal

Periodo: 2010-presente

FASE 3 BIG DATA

Sistema de extracción de datos mediante sensores móviles

1. Análisis centrado en la ubicación
2. Análisis centrado en la persona
3. Análisis de contexto
4. Visualización móvil

Figura 12. Las tres grandes fases de la historia del Big Data. Información adaptada de (Enterprise Big Data Framework, 2019).

Los Puntos Clave a través del Tiempo para el Big Data

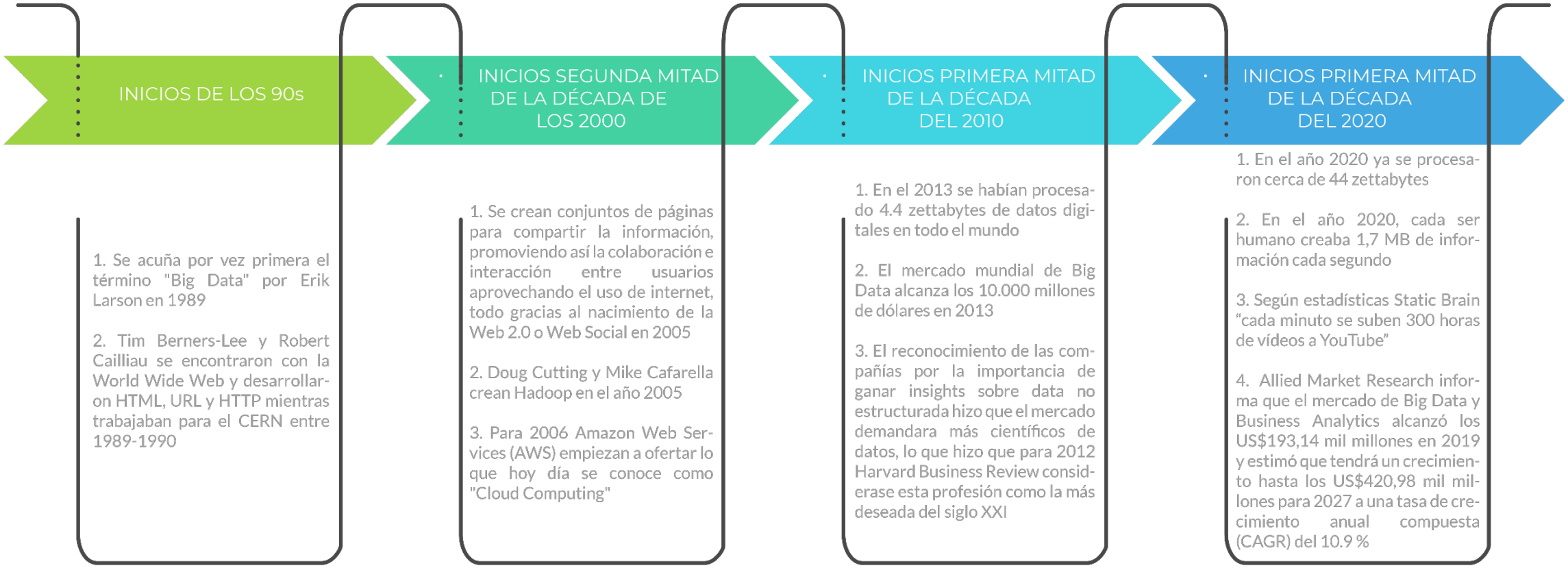


Figura 13. Los puntos clave a través del tiempo para Big Data.
Información adaptada de (Dynamic, 2020; Phillips, 2021).

Tendencias Actuales

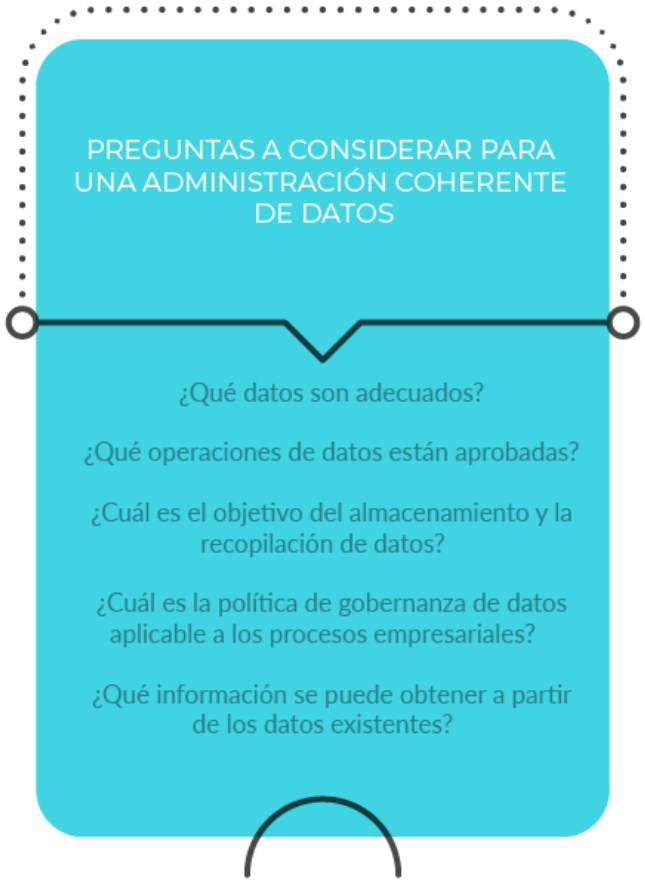


Figura 14. Preguntas a considerar para una administración coherente de datos. Información adaptada de (AWS, 2022).

- **Data Strategy:** Integrada para el crecimiento de las empresas. Esta estrategia busca alinear la visión a largo plazo de la organización en cuestión hacia un resultado que considere la recopilación, almacenamiento, intercambio y uso de los datos. Así pues, se pretenden obtener beneficios propios de un manejo óptimo de los datos como lo es un aumento en eficacia operativa, optimizar procesos, mejorar la toma de decisiones y directa o indirectamente mejorar tanto el nivel de servicio como el nivel de ingresos.

Tendencias Actuales

1. Resolución de desafíos planteados por la administración de datos

2. Mejora de la experiencia del cliente

3. Alcanzar la madurez analítica de acuerdo con el modelo de Gartner Analytic Ascendancy

4. Cree una cultura de datos que abarque a toda la organización

5. Cumplir con las regulaciones

Figura 15. Ventajas de aplicación de una estrategia de datos. Información adaptada de (AWS, 2022).

Tendencias Actuales

- **Tecnologías de la industria 4.0 en interacción:** Inteligencia artificial que implica almacenamiento masivo de datos a un menor costo, así como el uso hardware de alto rendimiento integrado al software. Esto llevado a aplicaciones como Ciudades Inteligentes (Smart Cities) y el IoT (Internet de las Cosas).

  En términos de la inteligencia artificial contamos con dos enfoques principales, los cuales son el enfoque humano y el enfoque más "ideal", donde el primero es uno donde los sistemas piensan como humanos y actúan como tal, mientras que el segundo es cuando los sistemas piensan racionalmente y actúan de tal manera (IBM Cloud Education, 2020).

  Aunque, no solo esta tecnología ha hecho que las interacciones humano-máquina cambie pues la amplia forma en que se pueden recabar datos ha llevado a que lleguemos a entender que la aceleración tecnológica ha hecho que las interfaces hombre-máquina (HMI) sean más sofisticadas al punto en que "las HMI utilizan sensores, robots, software, sistemas inalámbricos, software empresarial, aprendizaje de máquina a máquina u otras tecnologías para recopilar y analizar datos" (Rossi, 2016), mejorando así la calidad de las interacciones del humano con la tecnología y sus aplicaciones.

Gobernanza de Datos

BDPC™ Versión 092022

Definición

La definición de gobernanza de datos se puede dividir entre varias definiciones por distintas entidades o personas que han tenido impacto en el área; sin embargo, la que se utilizará será la del Data Governance Institute (s.f.) que, yendo más allá de definirla como un ejercicio de toma de decisiones y autoridad para los temas relacionados con la data, la define como:

“Un sistema de derechos de decisión y responsabilidades para los procesos relacionados con la información, ejecutado de acuerdo con modelos acordados que describen quién puede tomar qué acciones con qué información y cuándo, bajo qué circunstancias, utilizando qué métodos.”

Principales Actividades del Gobierno de Datos

Figura 16. Principales actividades del gobierno de datos.

Principales Actividades del Gobierno de Datos

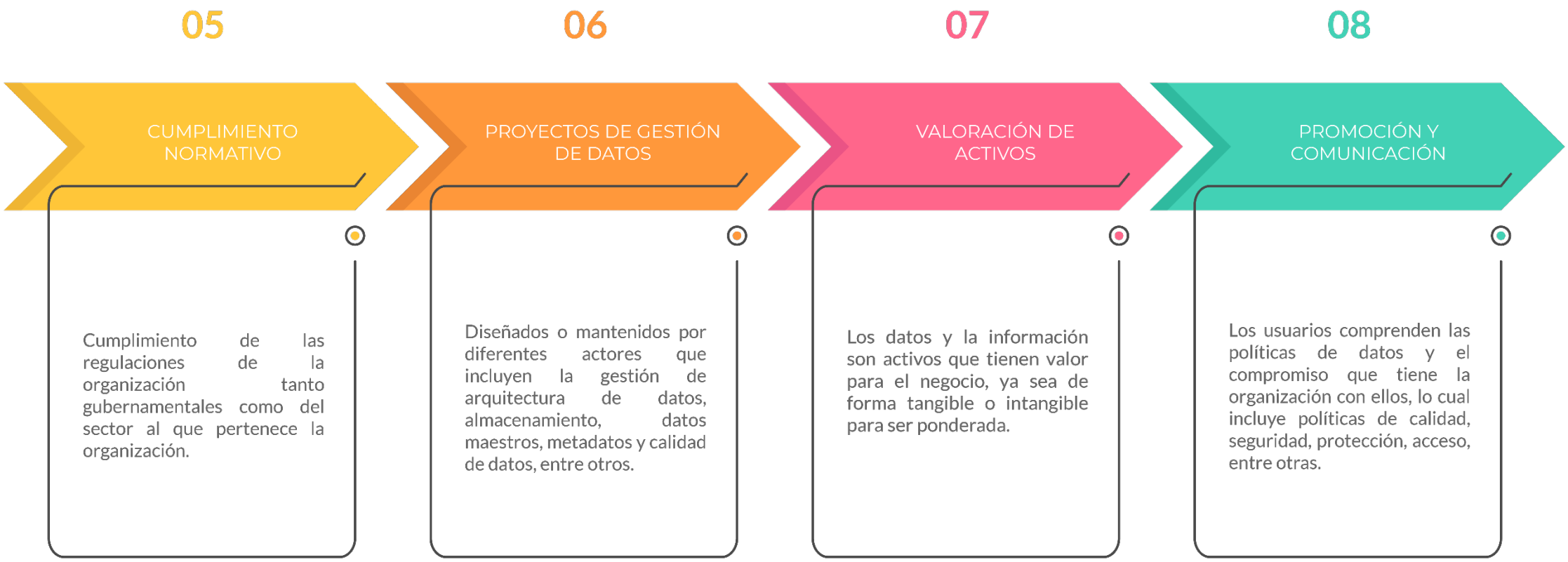


Figura 16. Principales actividades del gobierno de datos.

Ventajas del Gobierno de Datos

TOMA DE DECISIONES MÁS ACERTADAS Y OPORTUNAS

Se amplía la recepción de los datos necesarios para los usuarios de la organización, tanto a nivel interno como externo para así aprovechar las oportunidades de generar ingresos

MEJORA LOS CONTROLES DE COSTES

Al gestionarse los recursos de una manera más eficaz se pueden reducir problemáticas comunes tales como eliminar datos duplicados e innecesarios con mayor facilidad, o entender cuándo es viable o no adquirir hardware nuevo para la administración y almacenamiento de los datos

PERMITE QUE MÁS EMPLEADOS ACCEDAN A MÁS DATOS

Al contar con un alto rigor de gobierno de datos, se puede dar la permisibilidad de compartir el acceso a más datos ante los empleados ya que el factor seguridad y transparencia es más sólido

MEJORA EL CUMPLIMIENTO DE NORMATIVAS

La fijación rigurosa de las prácticas de gobernanza de datos permite a las organizaciones estar preparadas para alinearse a los marcos normativos que cada vez se tornan más complejos y estrictos frente a estas prácticas

GESTIÓN DE RIESGOS CON MÁS FACILIDAD

Se puede disipar las dudas sobre posibles exposiciones o malas prácticas sobre los datos al contar con un gobierno de datos riguroso, así como evitar intromisiones externas no deseadas

AUMENTA LA CONFIANZA QUE DEPOSITAN LOS CLIENTES Y PROVEEDORES EN LA ORGANIZACIÓN

La facilidad de demostrar por medio de auditorías el cumplimiento de las políticas de datos internas y externas hace que los stakeholders (partes interesadas) de la organización aumenten su confianza

Figura 17. Ventajas del gobierno de datos. Información adaptada de (Google Cloud, s.f.).

Roles en el Gobierno de Datos

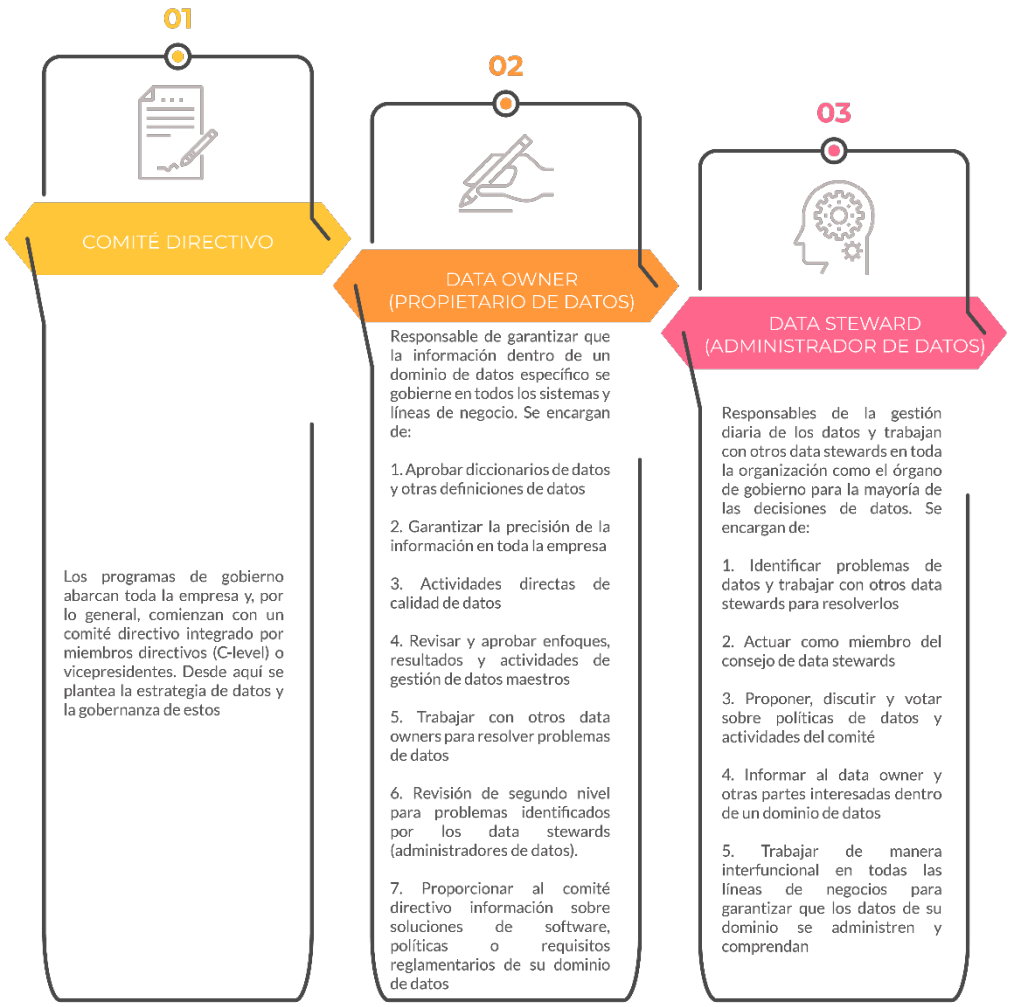


Figura 18. Roles en el gobierno de datos. Información adaptada de (Olavsrud, 2021).

Aplicaciones de Big Data

Aplicaciones de Big Data

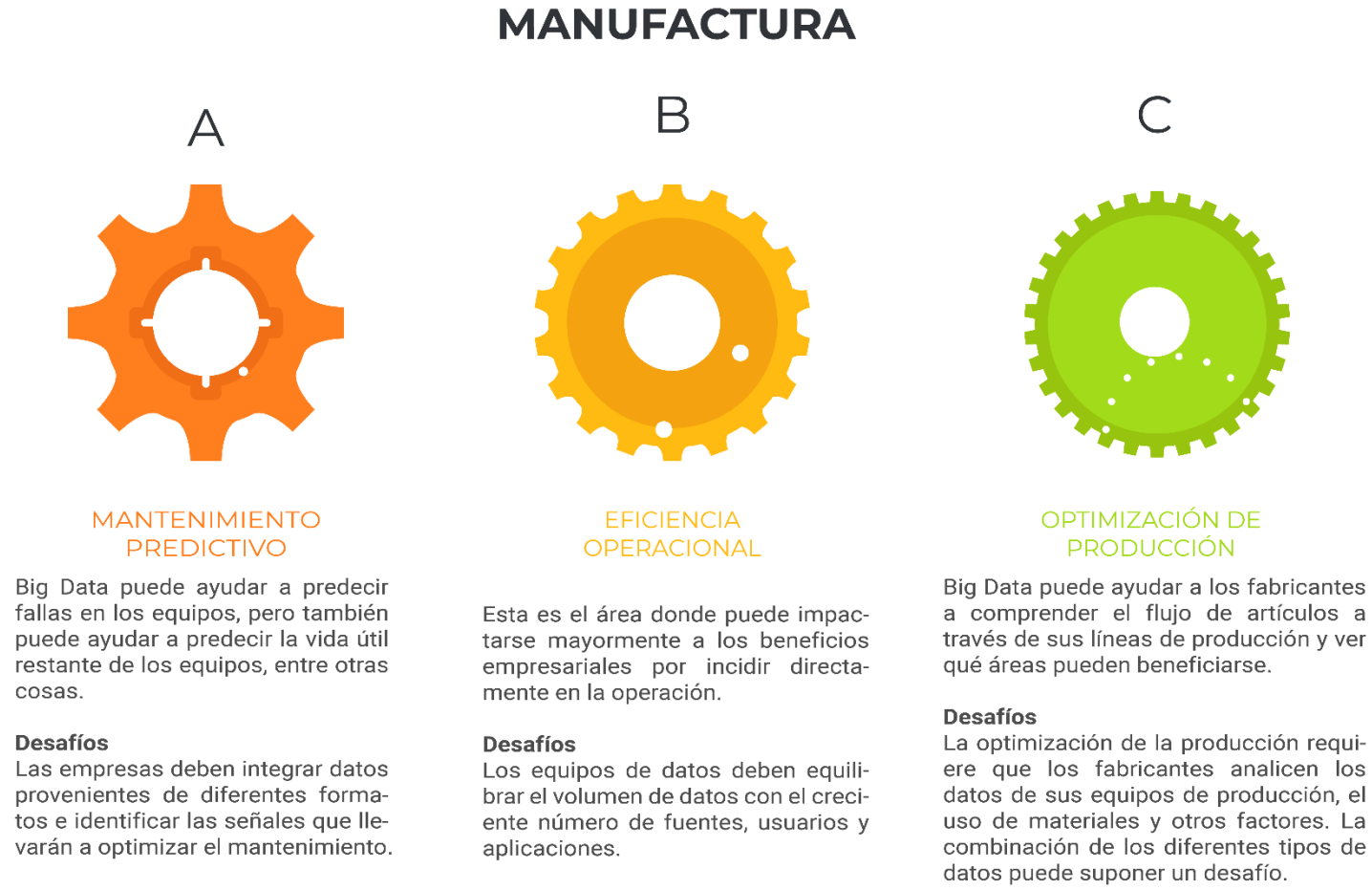


Figura 19. Desafíos para enfrentar en los casos de uso de Big Data Analytics de mayor impacto - Manufactura. Información adaptada de (Oracle, 2020).

Aplicaciones de Big Data

Figura 20. Desafíos para enfrentar en los casos de uso de Big Data Analytics de mayor impacto - Retail. Información adaptada de (Oracle, 2020).

Aplicaciones de Big Data

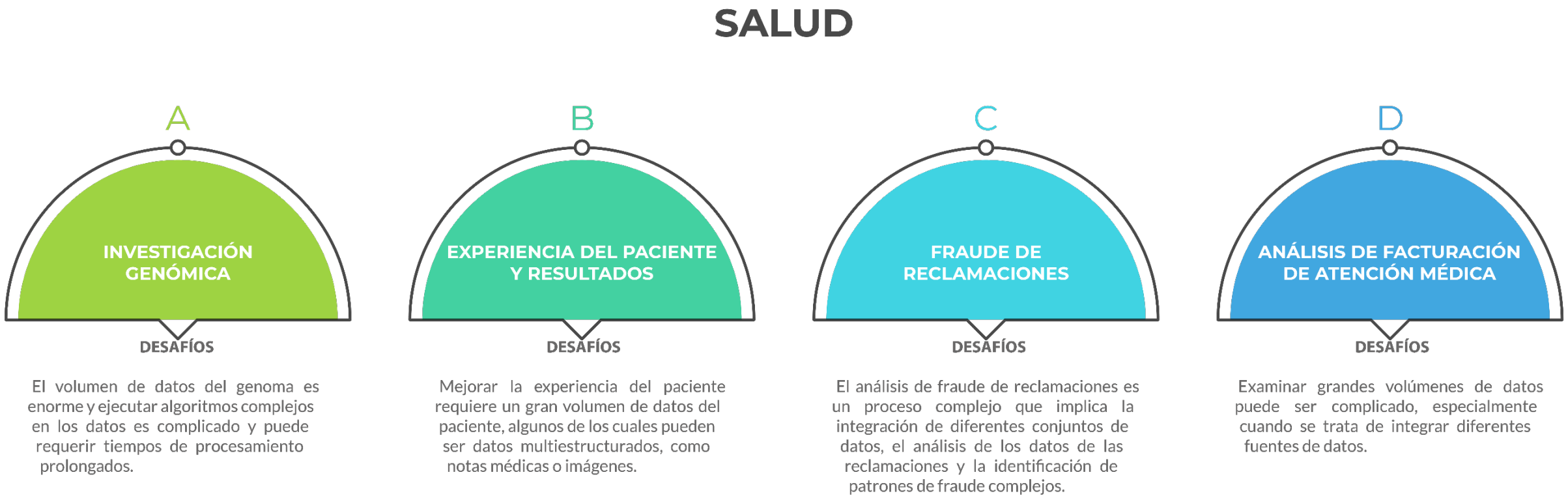


Figura 21. Desafíos para enfrentar en los casos de uso de Big Data Analytics de mayor impacto - Salud. Información adaptada de (Oracle, 2020).

Aplicaciones de Big Data

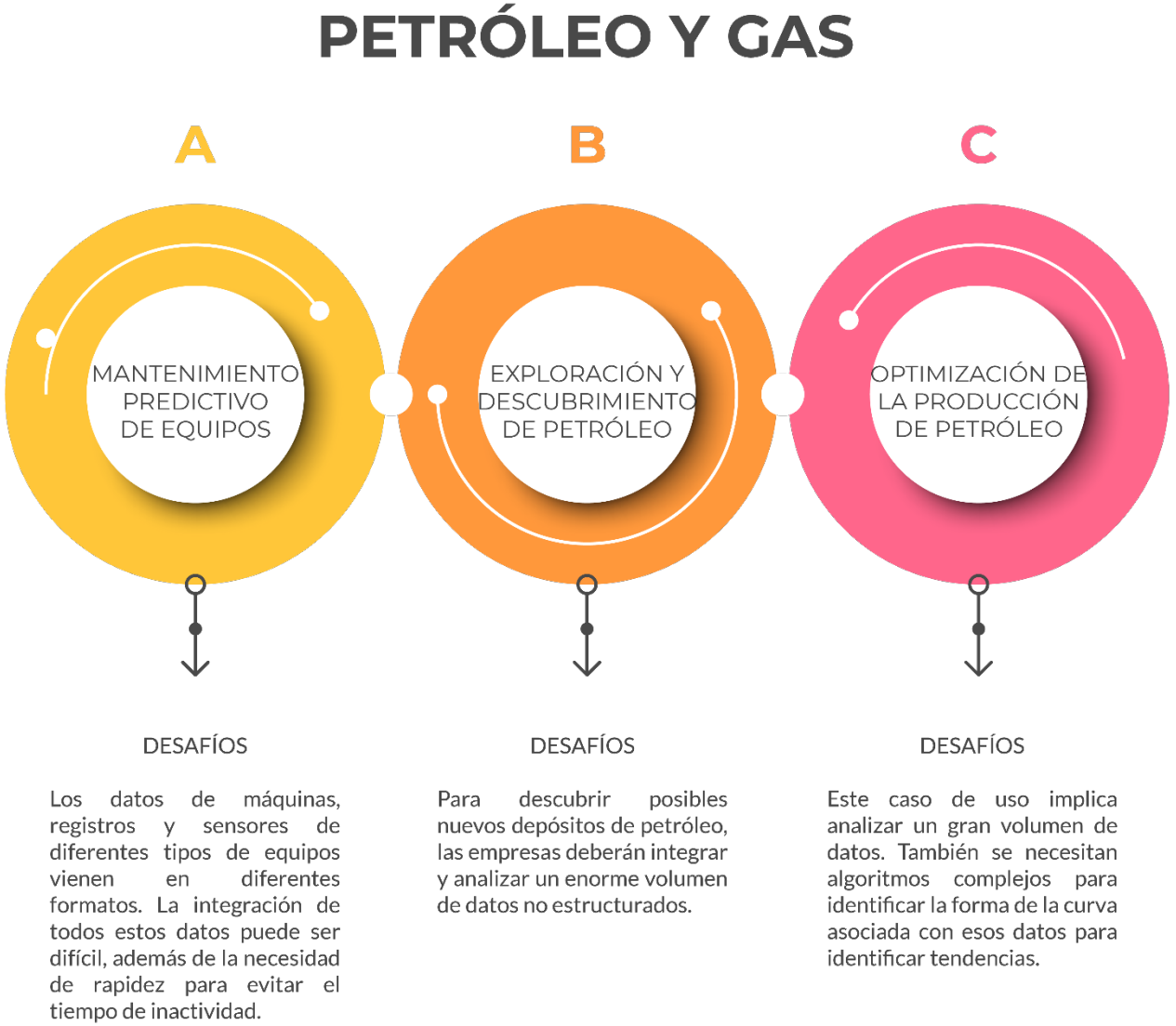


Figura 22. Desafíos para enfrentar en los casos de uso de Big Data Analytics de mayor impacto – Petróleo y gas. Información adaptada de (Oracle, 2020).

Aplicaciones de Big Data

Figura 23. Desafíos para enfrentar en los casos de uso de Big Data Analytics de mayor impacto – Telecomunicaciones. Información adaptada de (Oracle, 2020).

Aplicaciones de Big Data

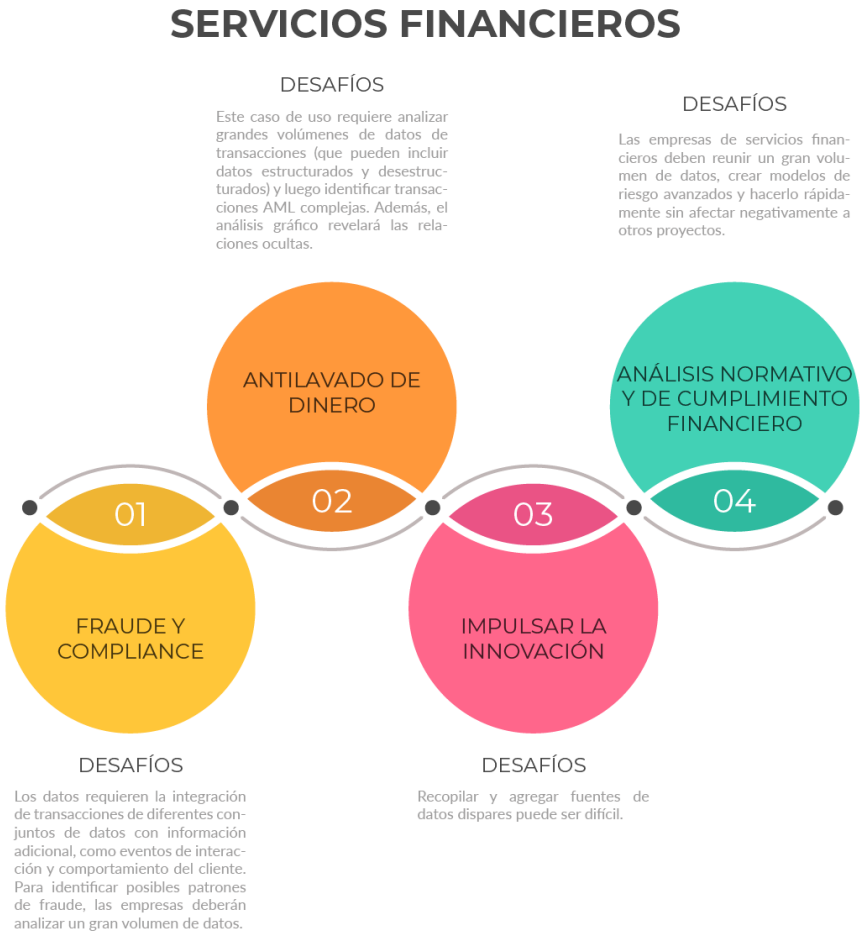


Figura 24. Desafíos para enfrentar en los casos de uso de Big Data Analytics de mayor impacto – Servicios financieros. Información adaptada de (Oracle, 2020).

Proyectos De Big Data con Business Intelligence (BI)

BDPC™ Versión 092022

¿Qué es el BI?

El Business Intelligence es un marco de trabajo que se alimenta de los datos de negocio para presentar una forma de visualización amigable con el usuario para así poder darle una manera más comprensible de entender qué es lo que está pasando con su negocio. Las herramientas que aportan a que se realicen prácticas de BI permiten acceder a datos de tipo histórico, actuales, de terceros, internos, entre otros (IBM, s.f.). Por lo tanto, BI permite proporcionar la información necesaria mediante el análisis de datos a las organizaciones para la mejora de la competitividad gracias a la toma de decisiones que va más allá de la extracción e informes de datos. (Castillo Romero, 2019). Para poder alcanzar esto es necesario considerar una serie de buenas prácticas a la hora de desempeñar labores de Big Data, así como entender en qué fases de la naturaleza de análisis se encuentra el BI.

¿Qué es el BI?

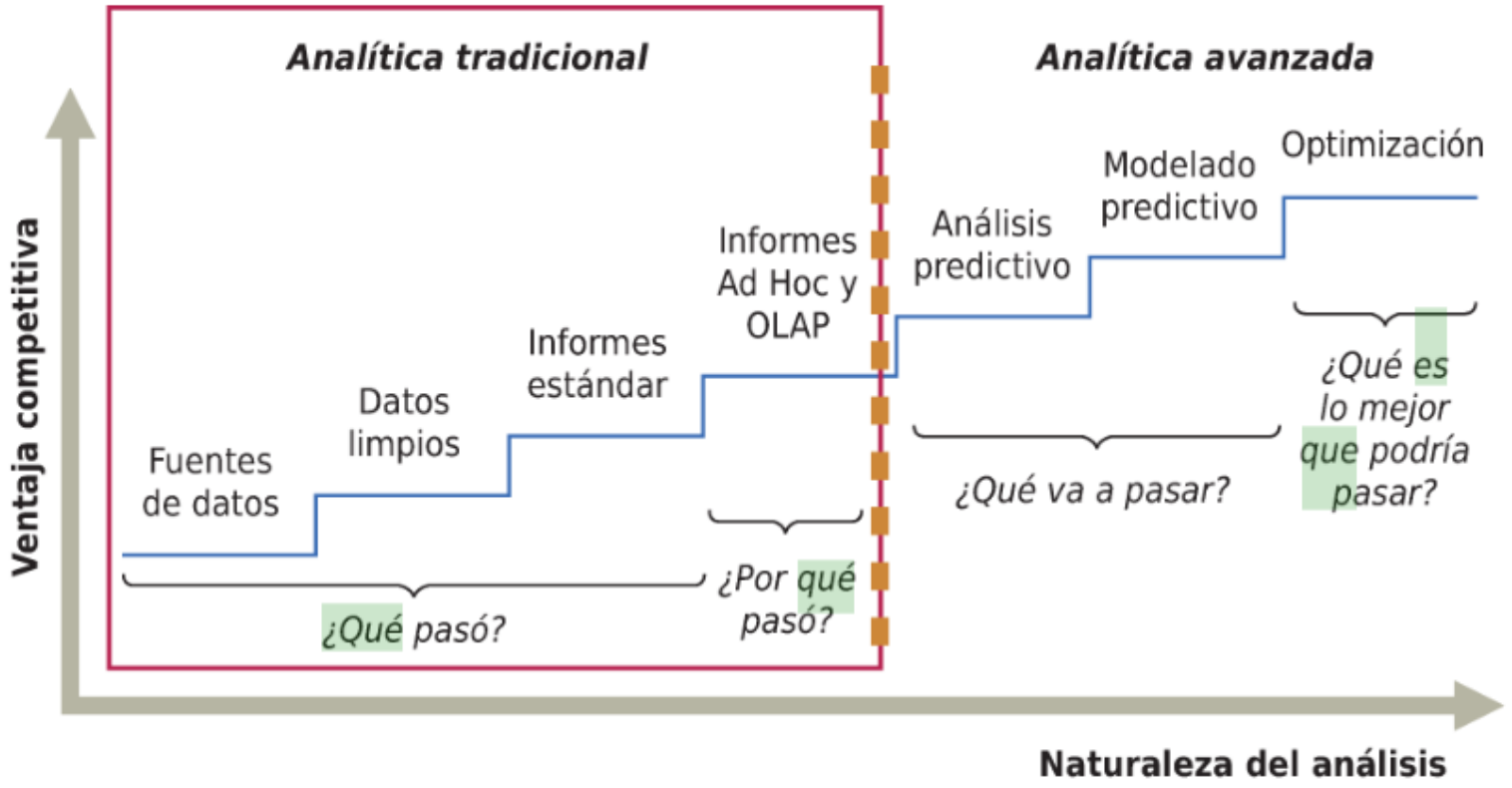


Figura 25. Analítica tradicional vs. Analítica avanzada. Obtenido de (Castillo Romero, 2019).

¿Qué es el BI?

Tradicionalmente, enfocarse en lo que está en la mitad de la izquierda de la gráfica es lo que hacen muchas organizaciones cuyo máximo alcance ha llegado hasta lo que son informes Ad Hoc que se encuentran hechos a manera de respuesta a una solicitud específica de información y lo que es OLAP (Procesamiento analítico en línea). Esta tecnología permite generar análisis multidimensional de data para así tomarla desde diferentes perspectivas y se suele organizar o estructurar en cubos de datos (un array de valores que permite representar y seguir data en más de 3 dimensiones) para finalmente almacenarse en data warehouses ya que métodos más tradicionales como hojas de cálculo de Excel se quedan cortos ya que es una lógica de dos dimensiones (una tabla de fila-columna). (Yvanovich, 2018).

¿Qué es el BI?

Sabiendo que OLAP es uno de los métodos que más soporta BI junto con otros métodos como consultas, informes y paneles, surge en un nivel superior lo que es el análisis predictivo que busca tratar de predecir eventos futuros con base en exploración de patrones de datos no procesados cuya detección puede ser de mayor complejidad, por lo que la analítica tradicional no termina de cumplir con alcanzar esas capacidades y así es que emerge como respuesta la analítica avanzada. Aunque, en este nivel de analítica avanzada hay dos escalones por subir como mínimo ya que lo que es el análisis predictivo como primer escalón es la respuesta a la pregunta de qué pasará, mientras que el siguiente escalón llamado análisis prescriptivo es la respuesta al qué debería pasar para trazar los caminos hacia dicha predicción. Finalmente, siempre se requerirá de un refinamiento u optimización para así lograr mejorar el resultado obtenido, lo cual siempre contará con distintas comparaciones y evaluaciones de los modelos planteados y entrenados.

¿Qué es el BI?

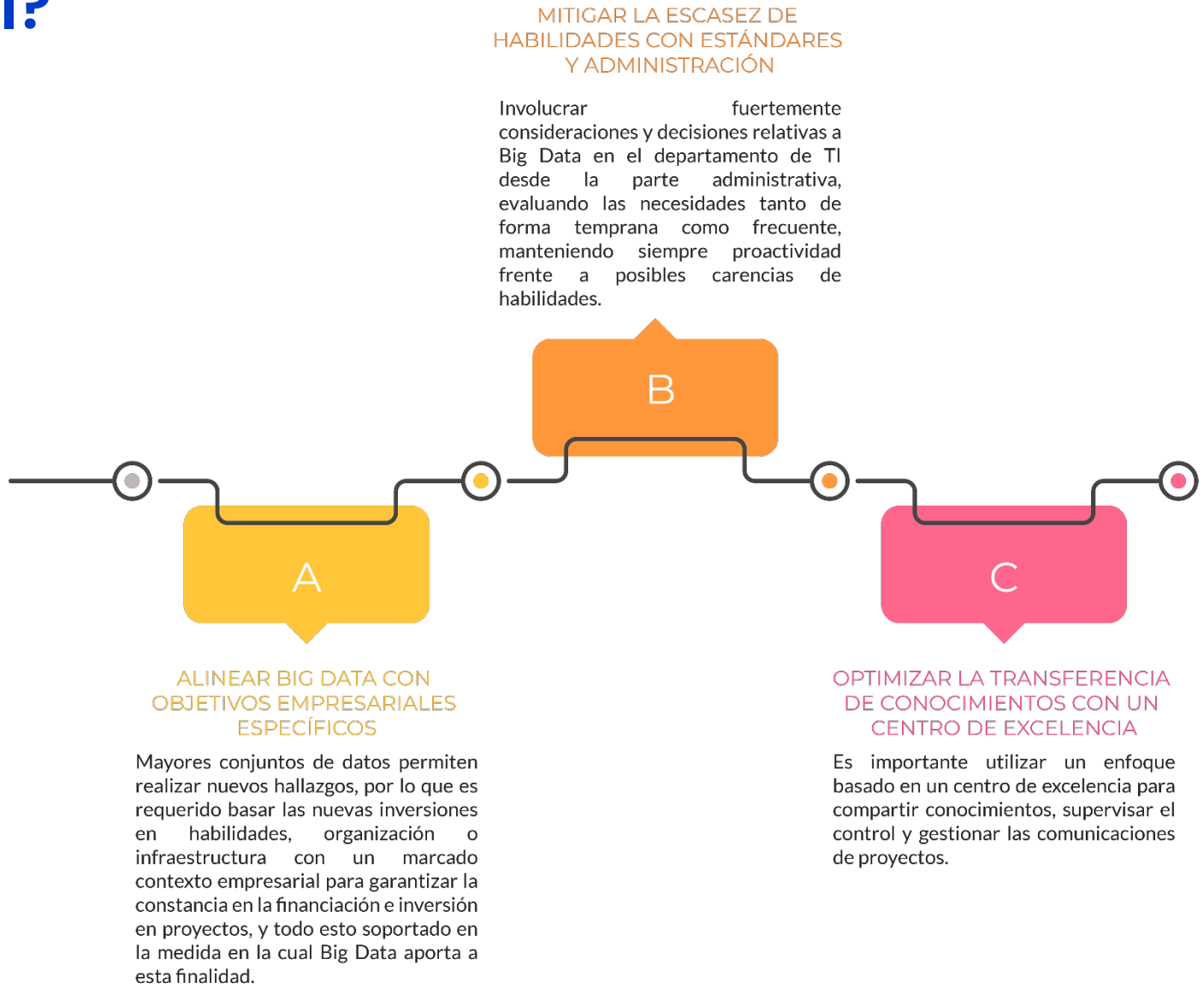


Figura 26. Mejores prácticas del Business Intelligence. Información adaptada de (Oracle, s.f.)

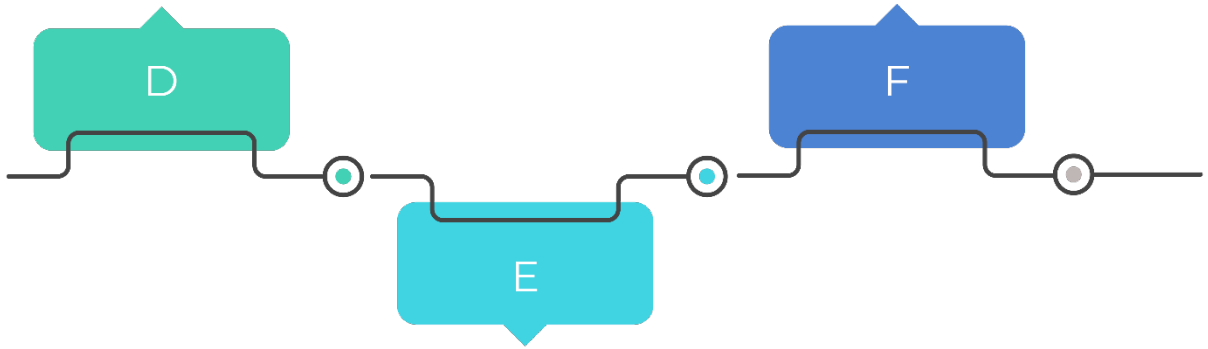
¿Qué es el BI?

LA PRINCIPAL VENTAJA RESIDE EN ALINEAR LOS DATOS ESTRUCTURADOS Y NO ESTRUCTURADOS

Integrar Big Data de baja densidad con datos estructurados de uso actual puede aumentar el valor del resultado final. Big Data debe considerar una extensión integral de las capacidades existentes de inteligencia empresarial (BI), de los sistemas de almacenamiento de datos y de las arquitecturas de información.

ALINEACIÓN CON EL MODELO OPERATIVO EN LA NUBE

Los entornos de pruebas (sandboxes) analíticos deben crearse on-demand. La gestión de recursos es fundamental para garantizar el control de todo el flujo de datos, incluido el procesamiento previo y posterior, la integración, el resumen dentro de la base de datos y la creación de modelos analíticos, así como mantener siempre seguridad frente al uso de nube pública y privada.



PLANIFICAR EL LABORATORIO DE HALLAZGOS EN PRO DEL RENDIMIENTO

El concepto "hallazgo" implica que los datos no siempre se obtienen directamente. En ocasiones, ni siquiera sabemos qué estamos buscando. La colaboración entre analísticas y científicos de datos es indispensable para entender las necesidades empresariales y sus carencias de conocimientos, manteniendo siempre entornos de prueba con el apoyo necesario y con una gobernanza propicia.

Figura 26. Mejores prácticas del Business Intelligence. Información adaptada de (Oracle, s.f.)

Métodos de BI

Figura 27. Métodos de Business Intelligence. Información adaptada de (Tableau, s.f.).

Los Data Warehouses (Almacenes de Datos)

Los data warehouses están diseñados para el análisis de datos ya que el procesamiento analítico es llevado a cabo cuando hay datos que han sido previamente preparados para que puedan pasar a fase de análisis para así generar conocimiento (Oracle, s.f.). Formados por grandes volúmenes de datos, distribuidos en varias unidades o grillas, ya sea por la gran cantidad de datos existente o por motivos de flexibilidad en la búsqueda de información. Además, permiten evitar posibles fallos que afectarían a toda la base de datos, así como ser considerados de las mejores opciones en cuanto a realizar tareas de analítica avanzada o análisis de datos históricos de fuentes variadas de datos.

Los Data Warehouses (Almacenes de Datos)

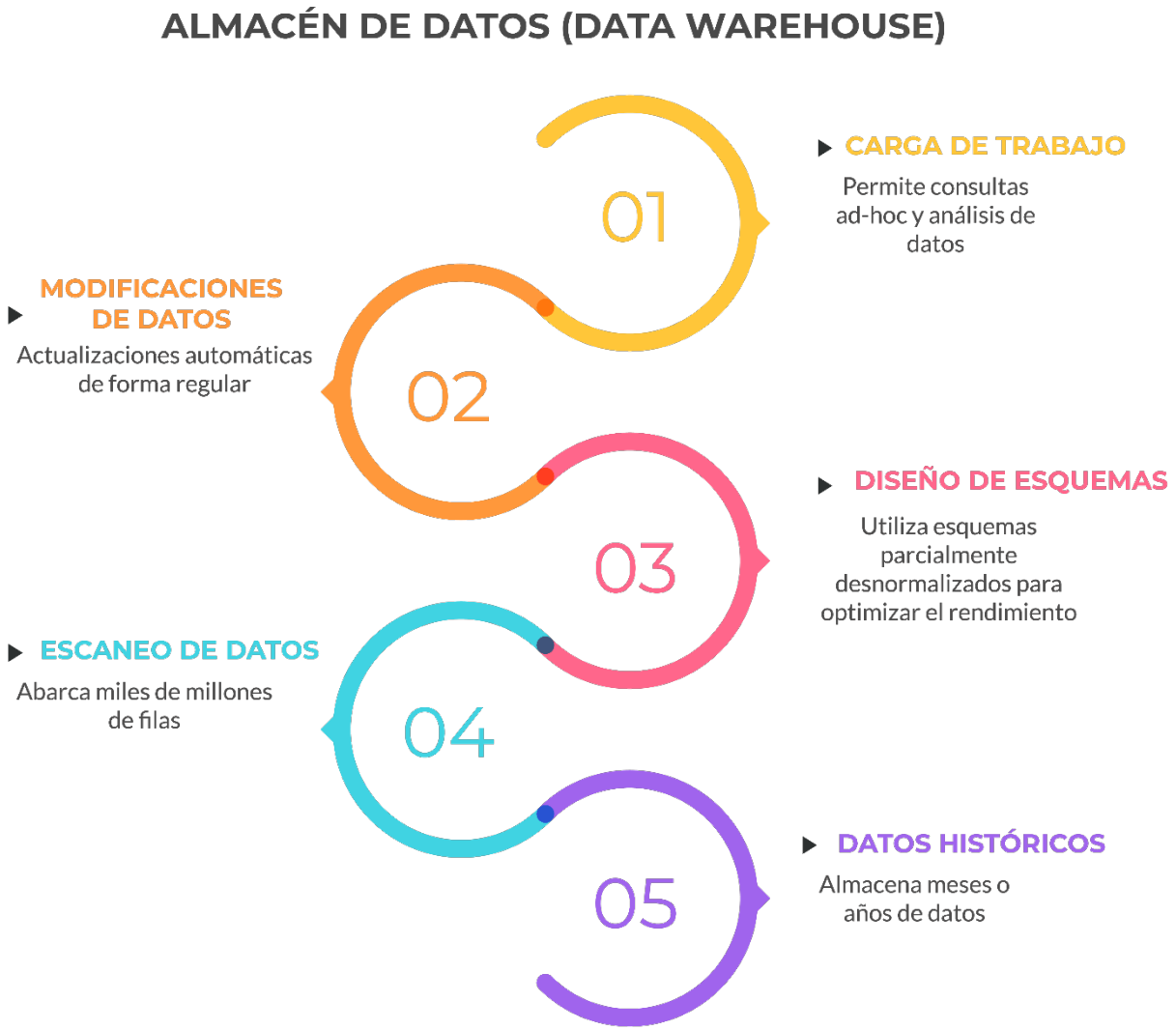


Figura 28. Características de un almacén de datos. Información adaptada de (Oracle, s.f.).

Evolución del BI y Analytics Frente a Orientación y Técnicas

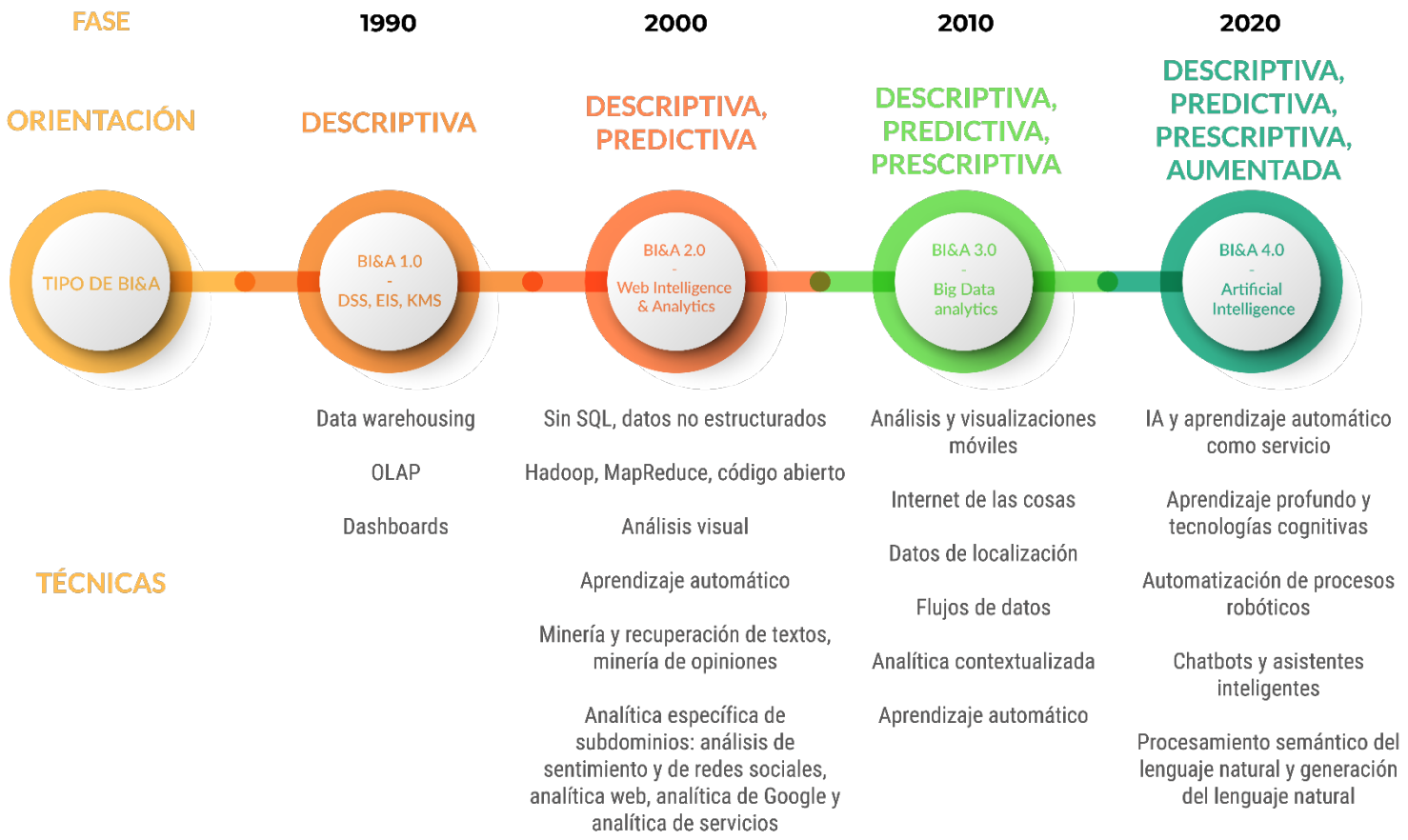


Figura 29. Evolución de Business Intelligence & Analytics (BI&A). Información adaptada de (Andoh-Baidoo et al., 2022).

La Metodología CRISP-DM. Big Data para la Minería de Datos

BDPC™ Versión 092022

Metodología CRISP-DM

La metodología CRISP-DM es una de las metodologías más completas por el hecho de que considera la aplicación al entorno de negocio donde se darán los resultados y lo que proporciona es "una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software" (Singular, s.f.). Se enfoca entonces en el ciclo de vida para un proyecto de minería de datos predictivo que permite obtener las variables estadísticas y predictivas que permitan generar un modelo de estudio (Bolbolian Ghalibaf, 2020).

Esta metodología contempla un contexto mucho más rico para la elaboración de los modelos y no descuida el hecho de que una vez finalizado igual hay momentos de despliegue y mantenimiento donde hay un acompañamiento constante al proyecto, así como la posibilidad de relación con otros proyectos. De manera que, es preciso mantener en todo momento una documentación precisa para los equipos desarrollo. Por otro lado, es importante entender que, si bien tiene una estructura cíclica, no es una secuencia de fases rígida pues se permite movilidad hacia delante y hacia atrás entre las diferentes fases (Singular, s.f.) de la metodología ya que una vez finalizada la labor de una fase se sabrá hacia dónde es más crítico moverse.

Metodología CRISP-DM

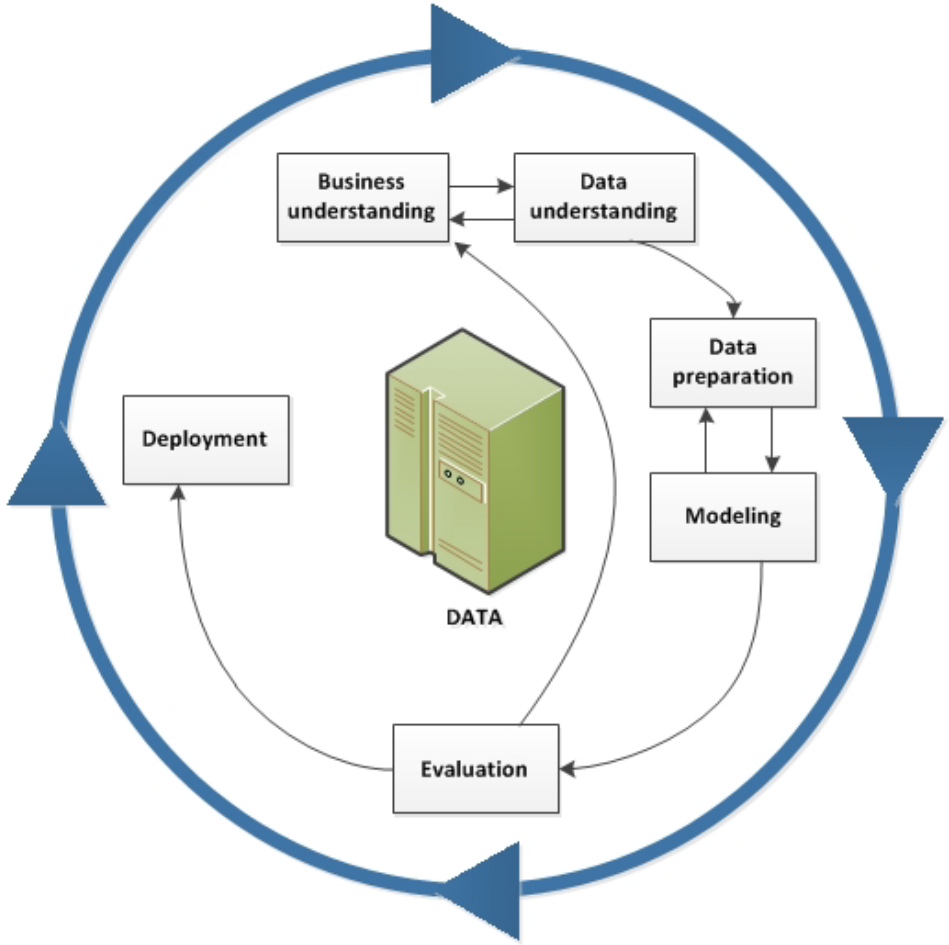


Figura 30. Ciclo de vida de minería de datos - Metodología CRISP-DM. Obtenido de (IBM, 2021).

Infografía

BUSINESS UNDERSTANDING

Definición de necesidades del cliente (comprensión del negocio)

Comprender los objetivos del proyecto y definición del problema de minería de datos

DATA UNDERSTANDING

Estudio y comprensión de los datos

Colección y familiarización con los datos para identificar problemas de calidad y descubrir conocimiento preliminar sobre estos datos, así como formar hipótesis

DATA PREPARATION

Análisis de los datos y selección de característi-cas

Construcción del conjunto final de datos a partir de los datos en bruto iniciales

MODELING

Modelado

Selección y aplicación de técnicas de modelado que sean pertinentes al problema, así como su calibración de parámetros a valores óptimos

EVALUATION

Evaluación (obtención de resultados)

Evaluar a fondo los modelos que han sido construidos y revisar los pasos ejecutados

DEPLOYMENT

Despliegue (puesta en producción)

Siempre existirá la necesidad de organizar y presentar lo que se haya realizado, y dependiendo de los requisitos puede variar la complejidad de esta fase

Figura 31. Fases de la metodología CRISP-DM. Información adaptada de (Singular, s.f.).

Metodología CRISP-DM como Evolución de la Metodología SEMMA

La metodología SEMMA es una que orienta su análisis y comprensión del problema hacia un aspecto más técnico y esto es dado porque desde su origen se planteó para trabajar con un software de explotación de datos. Esta corresponde a 5 fases distintas y comprende ser un conjunto de herramientas funcionales que se enfoquen hacia aspectos de desarrollo propio de un modelo de minería (Hernández G.& Dueñas R., 2009).

Metodología CRISP-DM como Evolución de la Metodología SEMMA

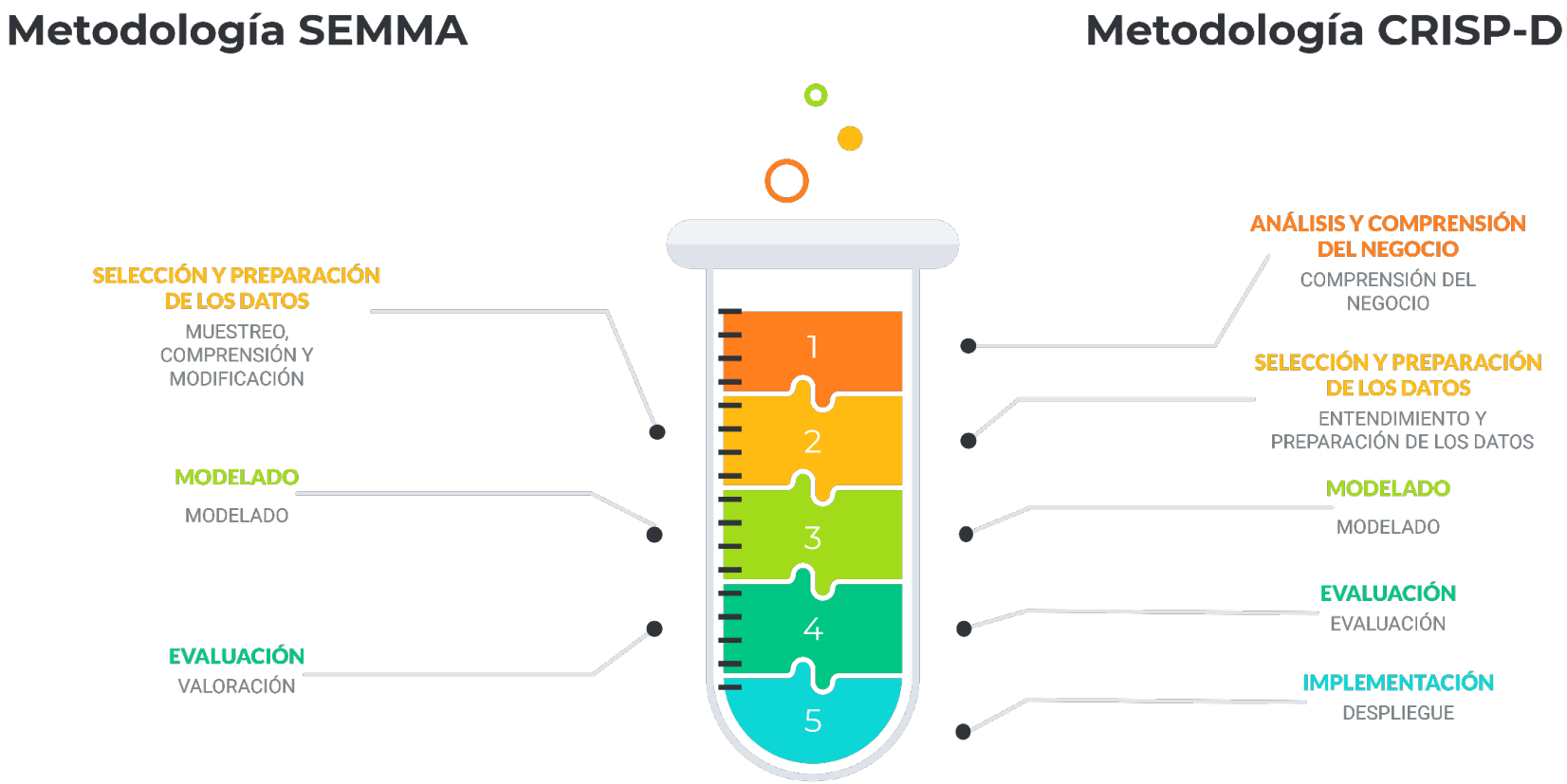


Figura 32. Comparativa de fases de las metodologías CRISP-DM y SEMMA. Información adaptada de (Peralta, 2014)

Metodología CRISP-DM como Evolución de la Metodología SEMMA

Las fases de la metodología SEMMA son como se describen a continuación: 1) Fase I – "Muestreo (Sample)" → Extracción de una muestra representativa; 2) Fase II – "Exploración (Explore)" → Exploración de los datos de la muestra seleccionada; 3) Fase III – "Modificación (Modify)" → Modificación de los datos; 4) Fase IV – "Modelado (Model)" → Modelación de los datos; y 5) Fase V – "Valoración (Assess)" → Evaluación de los datos. Estas fases dan una definición general de lo que se hará en la metodología, pero no suponen un detalle de tareas y actividades que se implementan dentro de la etapa, por lo que recae la responsabilidad en la organización que la implemente. Por otro lado, la metodología CRISP-DM sí provee esto, así como brindar actividades para la gestión del proyecto en cuanto a gestión del alcance, del tiempo, del costo, del riesgo y de los recursos humanos (Peralta, 2014).

Metodología CRISP-DM como Evolución de la Metodología SEMMA

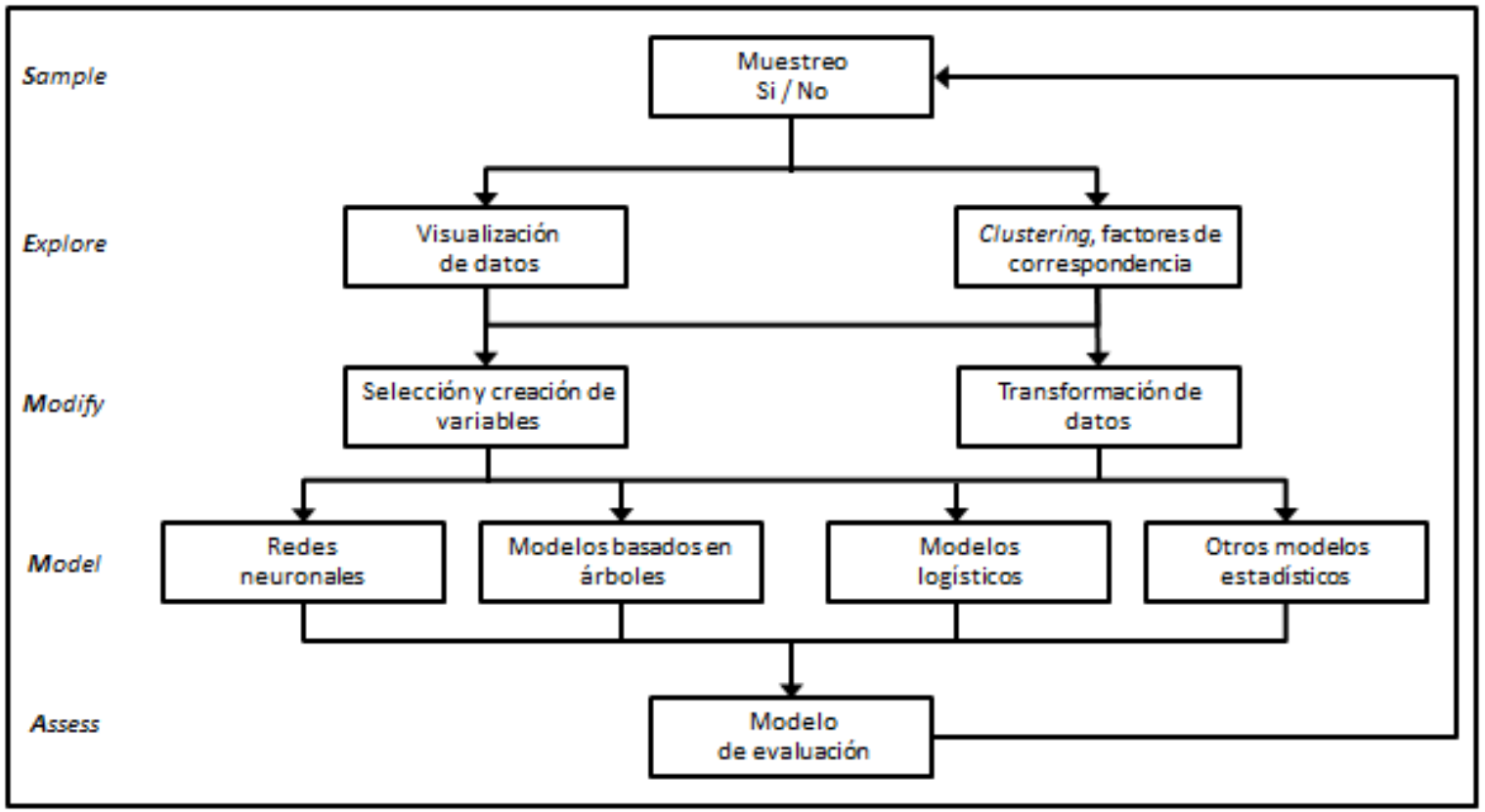


Figura 33. Dinámica general de la metodología SEMMA. Obtenido de (Peralta, 2014).

Análisis de Datos

BDPC™ Versión 092022

Análisis de Datos

De acuerdo con ISO (2019) en su norma ISO/IEC 20546:2019 se expone que el análisis de datos es utilizado "para comprender los objetos representados por los datos, hacer predicciones para una situación determinada y recomendar pasos para lograr los objetivos. Los conocimientos obtenidos de los análisis se utilizan para diversos fines, como la toma de decisiones, la investigación, el desarrollo sostenible, el diseño, la planificación, etc. Es un concepto compuesto que consiste en la adquisición de datos, la recopilación de datos, la validación de datos, el procesamiento de datos, incluida la cuantificación de datos, visualización de datos e interpretación de datos".

El análisis de datos aplicado a Big Data está actualmente asociado al aspecto competitivo de las organizaciones ya que puede mejorar la toma de decisiones y permitir extraer nuevos insights y conocimientos que antes de haber aplicado estas técnicas no habrían sido necesariamente posibles de conseguir, por ello la intrínseca relación entre Big Data y análisis de datos junto al desempeño de una organización (Ahmed et al., 2022).

Análisis de Datos

Además, la implementación de técnicas de Big Data al análisis de datos ha permitido encontrar nuevas formas de almacenar, analizar y visualizar datos de una manera avanzada y única para que así los insights generados sean más valiosos para las organizaciones, tanto a nivel de su operación, como a nivel de sus costos (Lin et al., 2022). Esto se ha visto reflejado en estudios donde se ha evidenciado que el análisis de datos con Big Data tiene generalmente un impacto positivo en lo que es la calidad de las decisiones organizacionales, así como las capacidades de análisis de datos a pesar de que el proceso para llegar a estos resultados no sea necesariamente sencillo (Li et al., 2022).

Tipos de Análisis

Figura 34. Tipos de análisis de datos.

Aprendizaje Supervisado

El aprendizaje supervisado es similar a cuando un estudiante está supervisado por un docente, se le está guiando hacia dónde debe dirigirse su resultado. Estos algoritmos permiten predecir fenómenos como lo pueden ser: pérdida de clientes, volumen de consumo, fraude, etc. n este caso, para la clasificación se fundamenta en la inferencia de ejemplos etiquetados para posteriormente hacer una predicción de un output para un conjunto de datos distinto donde los ejemplos no están etiquetados, siendo que esto es posible porque hay previamente un entrenamiento con las etiquetas.

A mayor entrenamiento, mayor exactitud en el rendimiento del algoritmo (sin llegar a un extremo de sobre entrenar el algoritmo como tal), y dependiendo de cuál sea la tarea por realizar es que un modelo es mejor que otro frente a sus mediciones de qué tan precisa es su predicción. Por ejemplo, se han dado estudios donde ha habido identificación de emociones por reconocimiento de imágenes de expresiones faciales con ayuda de un algoritmo de regresión logística (Barrionuevo et al., 2020).

Aprendizaje Supervisado

Se implementan los siguientes métodos:

MÉTODOS DE REGRESIÓN

1. Coeficientes de correlación (Correlación de Pearson)

2. Valores atípicos

3. Regresión Lineal: Permite predecir variables cuantitativas

4. Regresión Logística: Para predecir variables cualitativas binarias o multinomiales

5. Distribución normal de datos

6. Regresión de Poisson: Predicción de variables expresadas como porcentajes o conteos

7. Regresión de Cox: Para predicción duraciones

MÉTODOS DE ÁRBOL

1. Los árboles de decisión construyen particiones en los datos según el impacto de variables X (Explicativas), sobre la variable Y

2. Crear reglas de clasificación no ecuaciones

3. Algunos algoritmos son CHAID, CRT, Entropías

4.Algunas técnicas avanzadas de machine learning se basan en esta familia de técnicas, ejemplo: Random Forest, Gradient boosting

REDES NEURONALES

La red neuronal permite aprender sobre una variable a partir de otras disponibles, de la misma manera como el cerebro establece conexiones neuronales para generar reconocimiento y memoria sobre los datos. Ejemplos de su uso:

1. Realizar predicciones sobre un comportamiento de compradores a futuro

2. Automatización de actividades simples del marketing

3. Identificar segmentos de compradores por caracterización de patrones de compra

4. Realizar predicciones de ventas

Figura 35. Métodos de aprendizaje supervisado.

Aprendizaje No Supervisado

El aprendizaje no supervisado a diferencia del supervisado es que podría asemejarse a cuando una persona está teniendo una sesión autodidacta pues busca llegar a un resultado por su propia cuenta con los recursos que tiene a su alcance. Entonces, estos algoritmos permiten inferir patrones de un conjunto de datos sin una relación o referencia a resultados conocidos o etiquetados, por lo que facilitan agrupar datos no estructurados de acuerdo con sus similitudes y patrones diferentes sobre el conjunto de datos. El aprendizaje supervisado se puede dividir o agrupar en dos tipos, los cuales son el agrupamiento y la asociación, aunque es importante entender que cuando se está ejecutando un método de clustering donde hay datos etiquetados entonces deja de ser clustering y se torna en clasificación que es aprendizaje supervisado (Google Developers, 2022).

Aprendizaje No Supervisado

MÉTODO DE AGRUPAMIENTO

1. Exclusivo
2. Aglomerativo
3. Probabilístico
4. Solapamiento

MÉTODO DE ASOCIACIÓN

1. Establece asociaciones entre objetos de datos a partir de volúmenes de datos
2. Se crean clústeres que se dividen automáticamente en conjuntos de datos por grupos según sus similitudes
3. Producen instancias que forman el conjunto de datos por niveles de agrupación de tipo jerárquico
4. Se generan agrupamientos no jerárquicos como los k-means

Figura 36. Características de métodos de aprendizaje no supervisado.

Aprendizaje No Supervisado

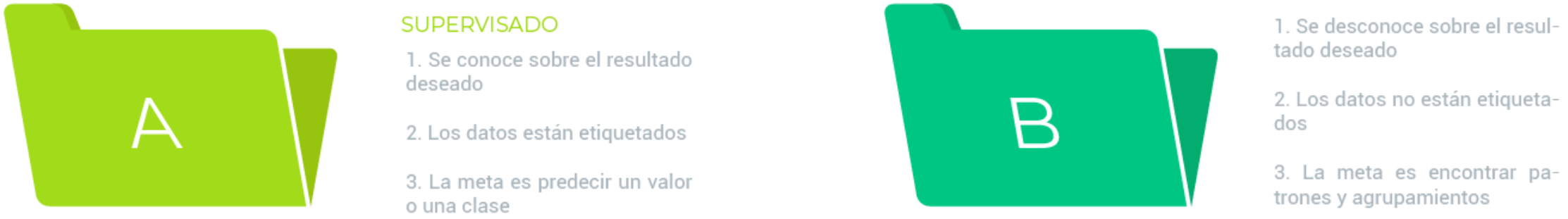


Figura 37. Diferencias entre aprendizaje supervisado y aprendizaje no supervisado.

Habilidades para el Análisis de Datos

Las habilidades necesarias para el análisis de datos con Big Data son las habilidades de liderazgo, técnicas, analíticas, estadísticas, de comunicación y de hacking ético. Estas habilidades son fundamentales para poder desempeñarse de la mejor manera en un ámbito que requiera de la realización de actividades relativas al análisis de datos, pero de una manera más avanzada al emplear técnicas de Big Data.

Habilidades para el Análisis de Datos

HABILIDADES ESTADÍSTICAS

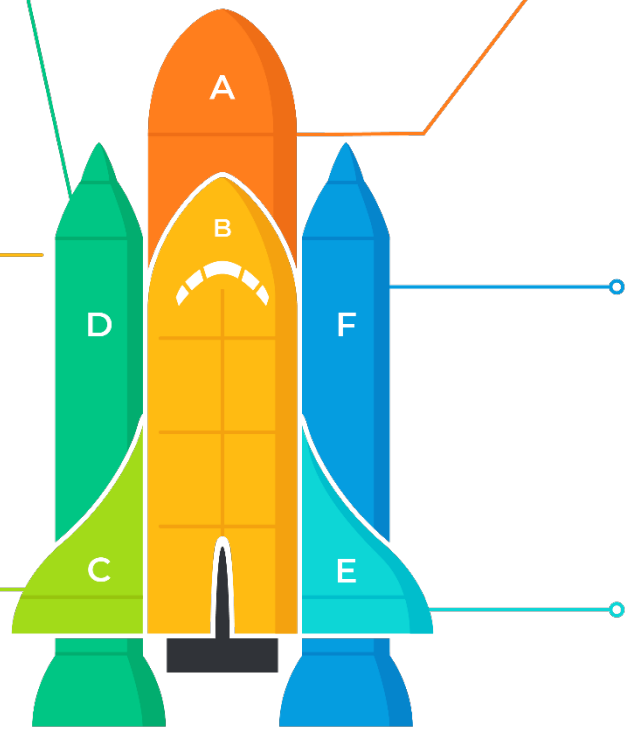
La comprensión sólida de los conceptos estadísticos básicos y sus operaciones subyacentes es indispensable para la evaluación y valoración de los modelos obtenidos a raíz de una labor de análisis de datos

HABILIDADES DE COMUNICACIÓN

Capacidad de traducir procesos y cálculos complejos en resúmenes visuales y fáciles de explicar a terceros que no hayan estado involucrados durante todo el proceso del análisis de datos, modelación y despliegue/implementación

HABILIDADES DE HACKING ÉTICO

Es la capacidad de encontrar distintas soluciones a los problemas que se afrontan, lo cual requiere de una profunda investigación con todos los recursos que estén a su disposición, pero sin transgredir en ningún momento la gobernanza, privacidad y tratamiento de los datos manejados



HABILIDADES DE LIDERAZGO

Capacidad de guiar/dirigir otros individuos dentro de la organización y la resolución de conflictos. Es indispensable que se prime una orientación hacia el análisis sobre los hechos o hallazgos en vez de juicios subjetivos que no estén alineados a lo que se desea obtener como resultado

HABILIDADES TÉCNICAS

Son las habilidades básicas en tecnologías que permitan un desempeño propicio para la elaboración de los modelos y las fases metodológicas necesarias para cumplir con la entrega de un resultado final basado en un análisis juicioso de los datos

HABILIDADES ANALÍTICAS

Es la capacidad de visualizar y analizar los datos, lo cual implica que haya una sinergia entre habilidades duras y blandas que permitan que las personas encargadas del análisis de los datos puedan brindar tanto juicios basados en su conocimiento técnico y del negocio, como en su creatividad, imaginación y curiosidad

Figura 38. Habilidades para el análisis de datos.

Tecnologías para Big Data

BDPC™ Versión 092022

Arquitectura de Referencia NIST

La arquitectura de Referencia de Big Data del NIST se define desde un enfoque neutral en cuanto a proveedores y puede ser utilizada por cualquier organización que pretenda implementar una solución en Big Data. Dentro de esta se encuentra un orquestador de sistema (system orchestrator), un proveedor de datos (data provider), un consumidor de datos (data consumer), un proveedor de aplicación Big Data (Big Data application provider), y un proveedor del marco Big Data (Big Data framework provider). Este último tiene a su vez marcos de procesamiento (processing frameworks), plataformas (platforms), infraestructuras (infrastructures) y recursos tanto físicos como virtuales (physical and virtual resources), además de que los 3 primeros componentes son escalables vertical y horizontalmente.

Por otro lado, el proveedor de aplicación Big Data es el que tiene las responsabilidades de recolección (collection), preparación (preparation), analítica (analytics), visualización (visualization) y acceso (access) a todos los datos manejados. Todos los componentes previamente mencionados hacen parte de una cadena de valor de información a nivel horizontal y de IT a nivel vertical, por lo que es necesario que todo se contenga de manera segura y privada (security and privacy fabric), así como manejar una administración (management fabric) adecuada de todo este marco de trabajo de Big Data.

Arquitectura de Referencia NIST

SYSTEM ORCHESTRATOR

El orquestador del sistema proporciona los requisitos generales que debe cumplir el sistema, incluidos los requisitos de política, gobierno, arquitectura, recursos y negocios, así como actividades de supervisión o auditoría para garantizar que el sistema cumpla con esos requisitos.

DATA PROVIDER

Crea una abstracción de varios tipos de fuentes de datos (como datos sin procesar o datos transformados previamente por otro sistema) y los pone a disposición a través de diferentes interfaces funcionales.

BIG DATA APPLICATION PROVIDER

El proveedor de aplicaciones de Big Data ejecuta las manipulaciones del ciclo de vida de los datos para cumplir con los requisitos establecidos por el orquestador del sistema.

BIG DATA FRAMEWORK PROVIDER

Tiene recursos o servicios generales para ser utilizados por el proveedor de aplicaciones de Big Data en la creación de la aplicación específica, y también consta de una o más instancias de los tres subcomponentes: marcos de infraestructura, plataformas de datos y marcos de procesamiento.

DATA CONSUMER

Las fábricas de seguridad y privacidad interactúan con el orquestador del sistema para políticas, requisitos y auditorías, y también con el proveedor de aplicación Big Data y el proveedor del marco de trabajo Big Data para desarrollo, implementación y operación.

MANAGEMENT FABRIC

Las características de Big Data de volumen, velocidad, variedad y variabilidad exigen una plataforma de administración de software y sistema versátil para el aprovisionamiento, la configuración y administración de software y paquetes, junto con la supervisión y administración de recursos y rendimiento.

Figura 39. Taxonomía de la arquitectura de referencia NIST. Información adaptada de (Chang et al., 2019)

Arquitectura de Referencia NIST

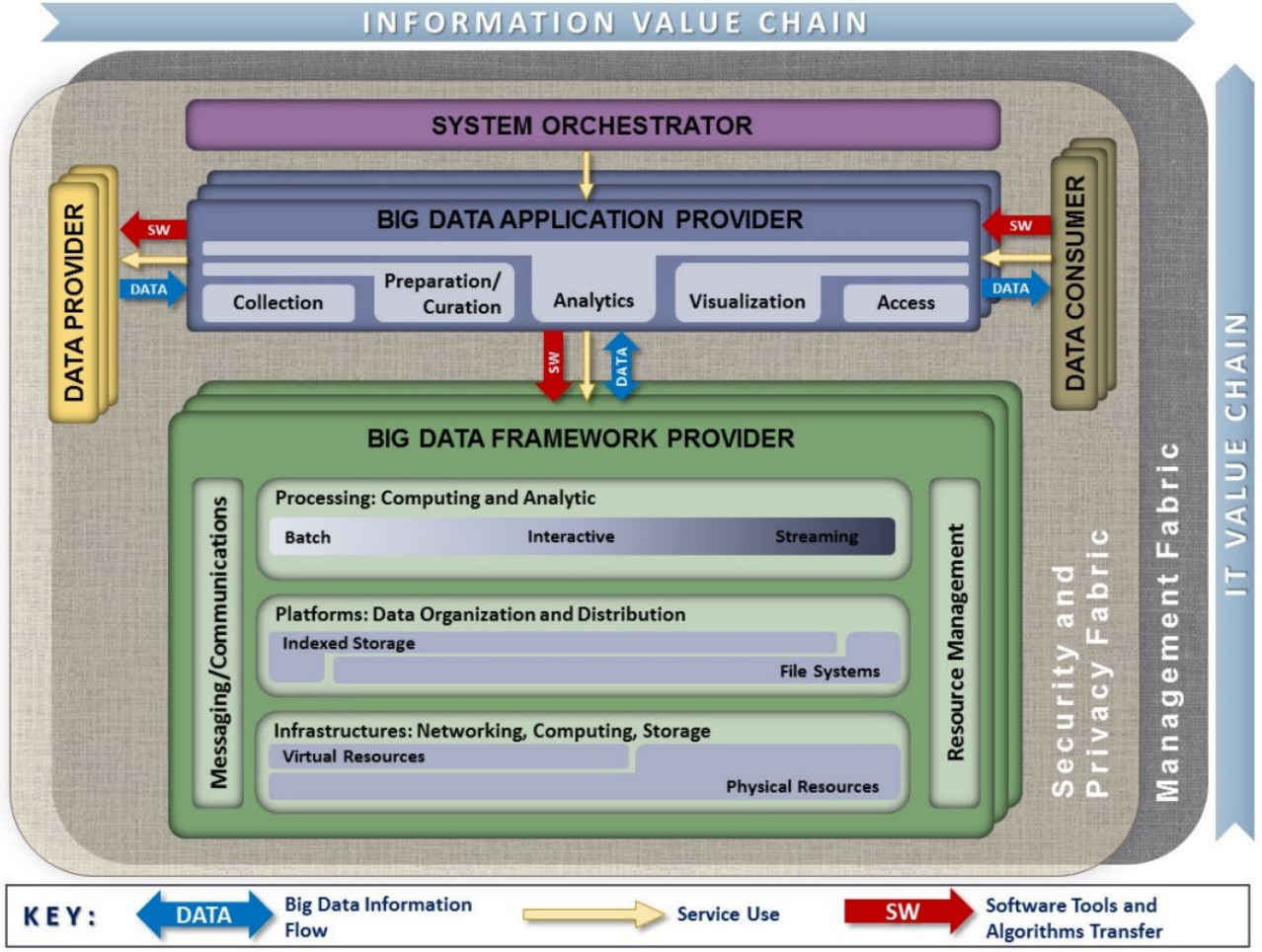


Figura 40. Arquitectura de referencia Big Data NIST (NBDRA). Información obtenida de (Chang et al., 2019).

Arquitectura Big Data Propuesta por Microsoft

Este modelo se desglosa en cuatro componentes que son: fuentes de datos (Data Sources), transformación de datos (Data Transformation), infraestructura de datos (Data Infrastructure) y uso de datos (Data Usage). Las fuentes de datos pueden ser variadas y pueden traer distintos tipos de datos con diferentes características, pero hay que tener presente que cumplen con las Vs de Big Data y su propósito puede llegar a cambiar o evolucionar conforme pase el tiempo ya que los datos recogidos pueden servir para diferentes fines.

Por otro lado, la transformación de datos consta de cuatro subetapas que son la recolección de datos, la agregación de nuevos datos, la congruencia entre todos los datos recogidos y sus metadatos, y la minería de datos para así hallar relación entre los datos. La infraestructura de Big Data es la que permite que se brinden los recursos necesarios para que todo funcione y el uso de datos es dependiente del usuario siempre y cuando se mantengan consideraciones de seguridad, así como su presentación en formatos variados (Vega et al., 2015).

Arquitectura Big Data Propuesta por Microsoft

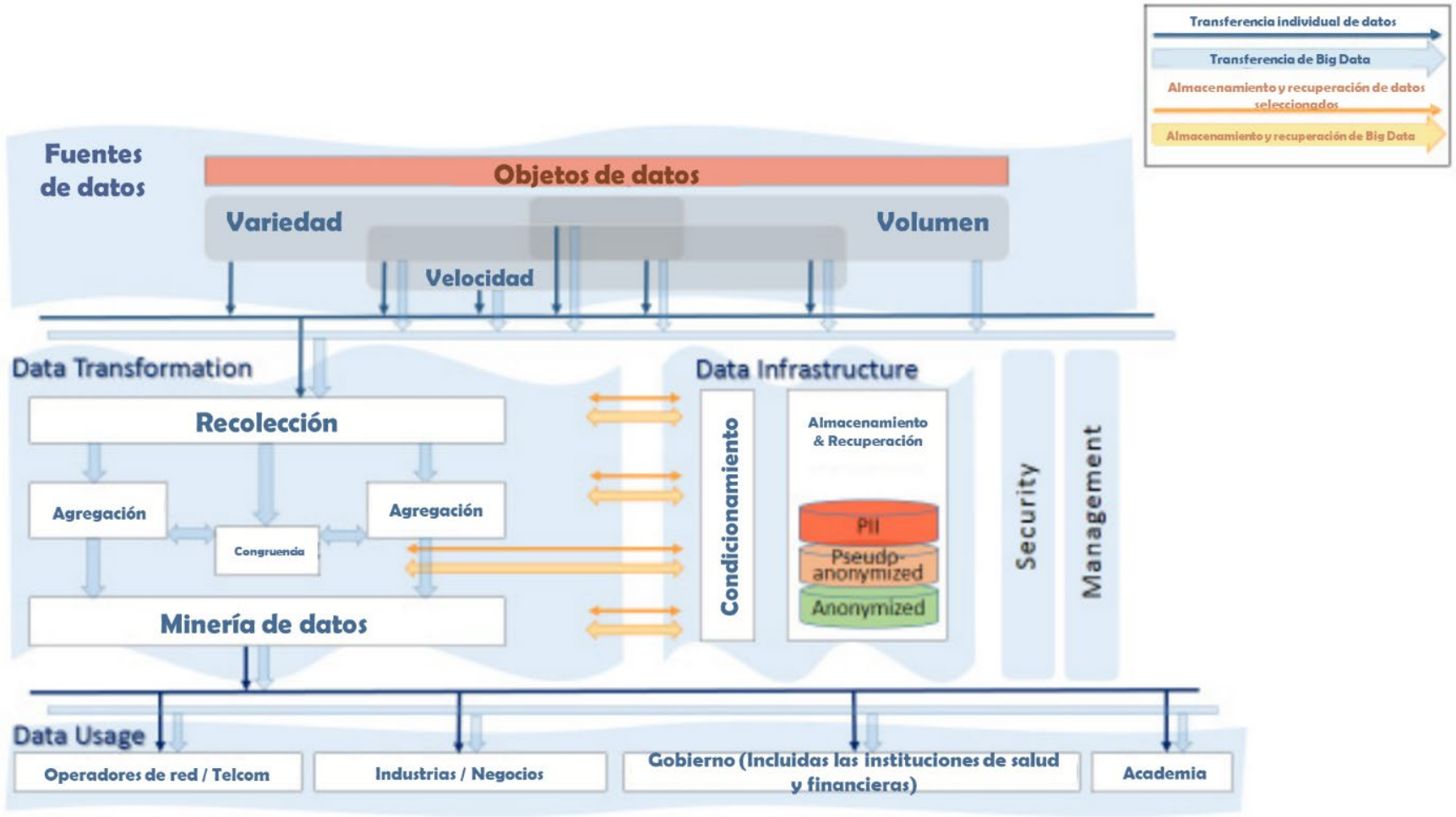


Figura 41. Arquitectura Big Data propuesta por Microsoft. Obtenido de (Vega et al., 2015).

Arquitectura MapReduce

En esta arquitectura hay trabajos/tareas que se rastrean en modo maestro-esclavo y hay dos tipos de trabajos, los cuales se conocen como mapeadores (los que mapean el trabajo en muchas tareas) y los reductores (los encargados de combinar las entradas de los mapeadores en una salida única). Para estos efectos, un proceso es maestro (JobTracker que es el que rastrea el contenido de los TaskTrackers) y el resto son esclavos (TaskTracker que es el que rastrea los procesos en un nodo computacional). Este tipo de implementaciones se asocian para el procesamiento y generación de grandes conjuntos de datos y son los usuarios los que especifican una función map que procesa un par clave/valor para generar un conjunto de pares intermedios clave/valor y una función reduce que fusiona todos los valores intermedios asociados a la misma clave intermedia.

Por otro lado, los programas escritos en este estilo funcional se paralelizan automáticamente y se ejecutan en un gran clúster de máquinas de consumo, asimismo, el sistema de ejecución se encarga de: los detalles de la partición de datos entrada, la programación de la ejecución del programa en un conjunto de máquinas, la gestión de los fallos de las máquinas y la gestión de la comunicación necesaria entre máquinas. Finalmente, MapReduce permite a los programadores sin experiencia en sistemas paralelos y distribuidos utilizar fácilmente los recursos de un gran sistema distribuido.

Arquitectura MapReduce

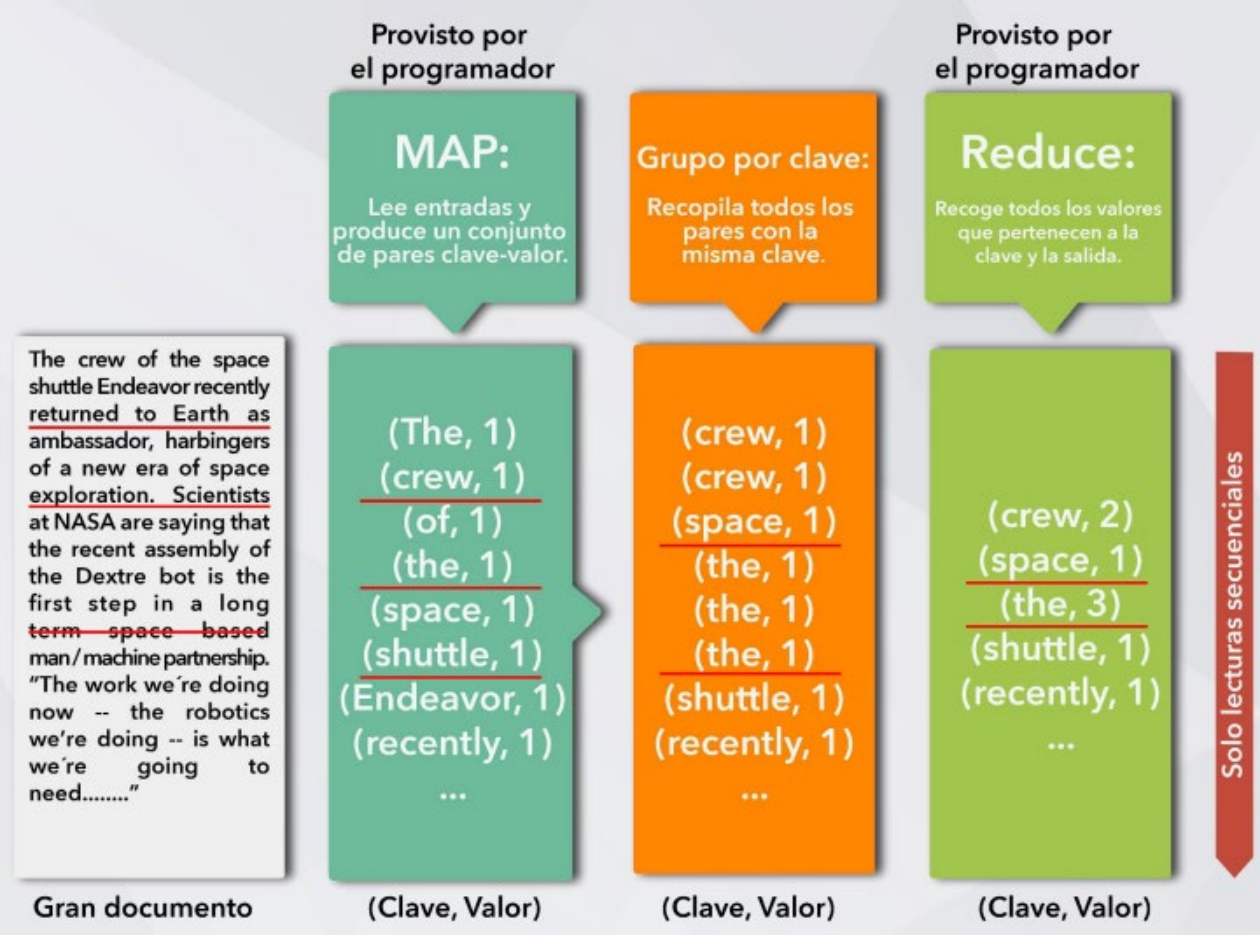


Figura 42. Contador de palabras usando MapReduce.

Arquitectura MapReduce

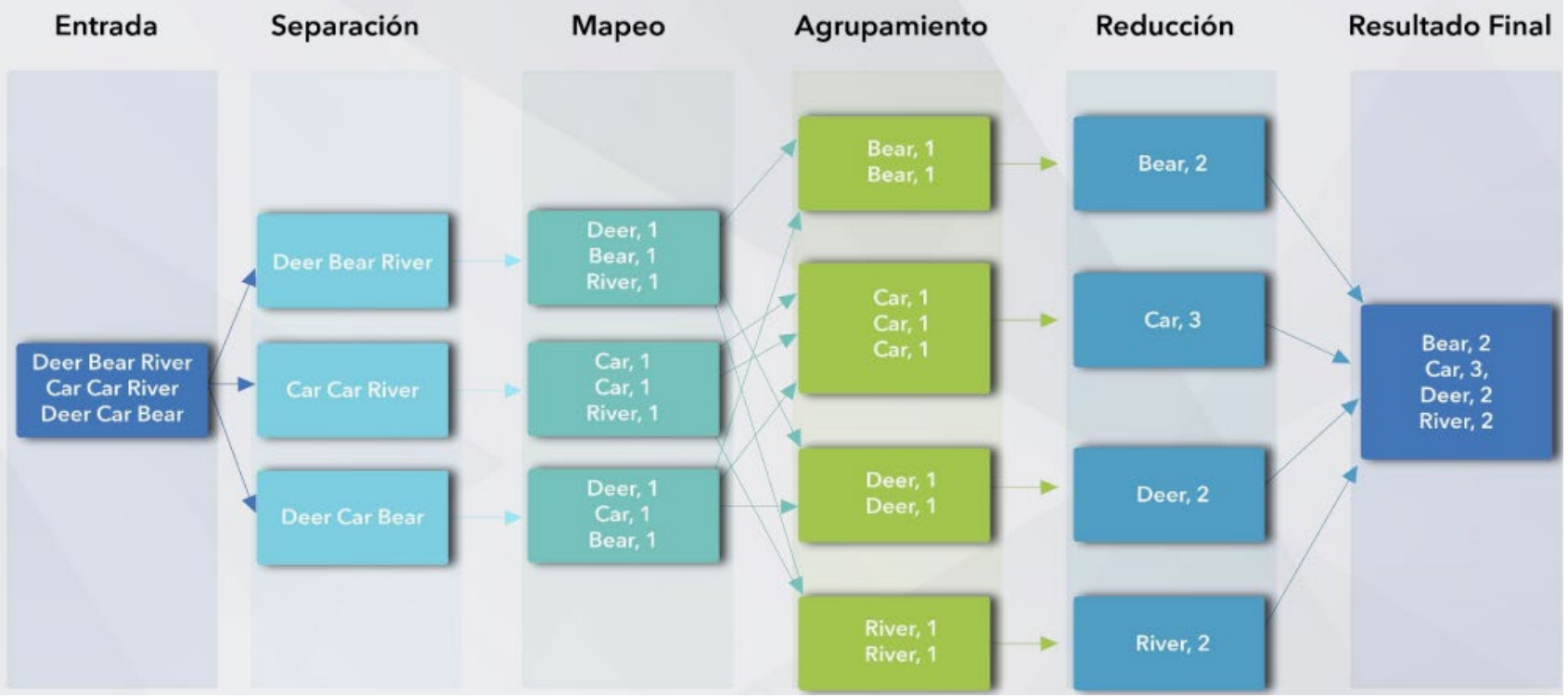


Figura 43. Proceso general de conteo de palabras de MapReduce.

Arquitectura NoSQL

Las bases de datos NoSQL son de diseño no relacional y son útiles para datos tanto de carácter estructurado como no estructurado, aunque es importante considerar que los archivos en estas bases de datos se escriben una vez y casi nunca se actualizan. A nivel de aplicaciones están principalmente diseñadas para aplicaciones descentralizadas (web, móvil, sensores, etc.) que cuentan con una disponibilidad continua para recibir y servir datos con alta velocidad de datos y baja latencia de acceso. En adición, son de bajo costo, varios tipos de herramientas de software son de código abierto, y su tecnología es de muchos diseños con muchas implementaciones de estructuras de datos e idiomas de acceso.

Sus esquemas suelen ser dinámicos ya que pueden añadirse nuevos campos según se requiera y pueden requerir reglas de validación de datos, además de ser altamente escalables a nivel horizontal para un aumento de capacidad en cuanto se añadan servidores básicos o instancias de la nube (MongoDB, s.f.). Finalmente, entre sus tipos está el valor clave (Amazon SimpleDB), la familia de columnas (Cassandra o HBase), las bases documentales (MongoDB) o las gráficas (Neo4J), y entre sus arquitecturas populares están: Master, Slave, HBase, Ring Architecture, Cassandra.

Arquitectura NoSQL

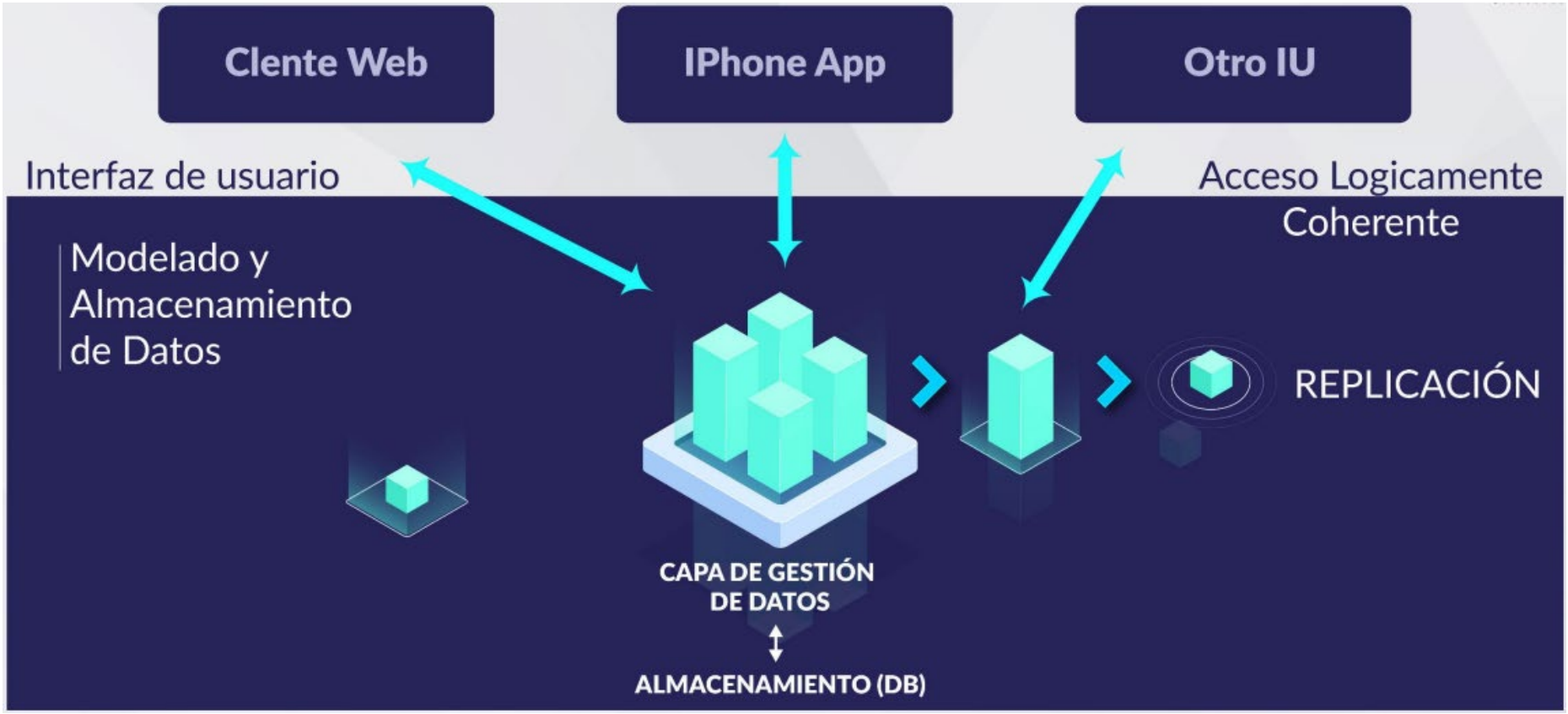


Figura 44. Arquitectura NoSQL.

Arquitectura NoSQL

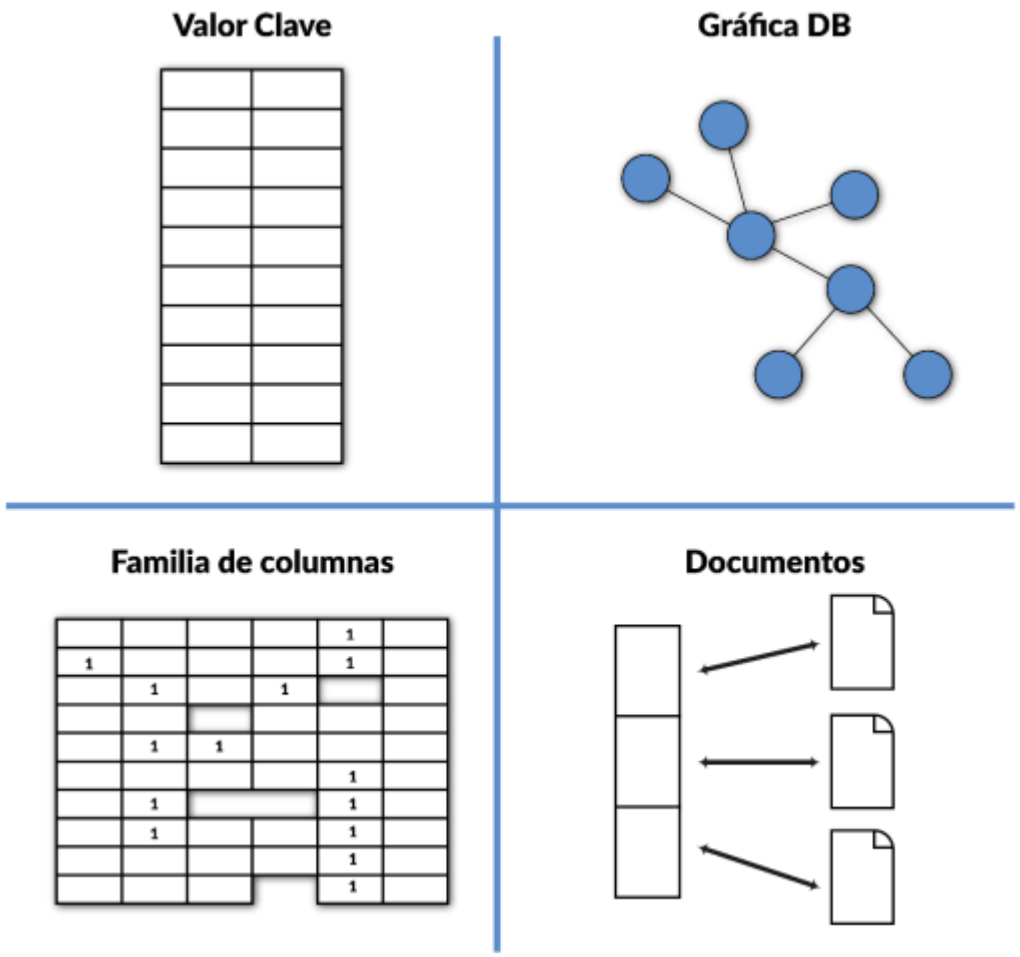


Figura 45. Tipos de bases NoSQL.

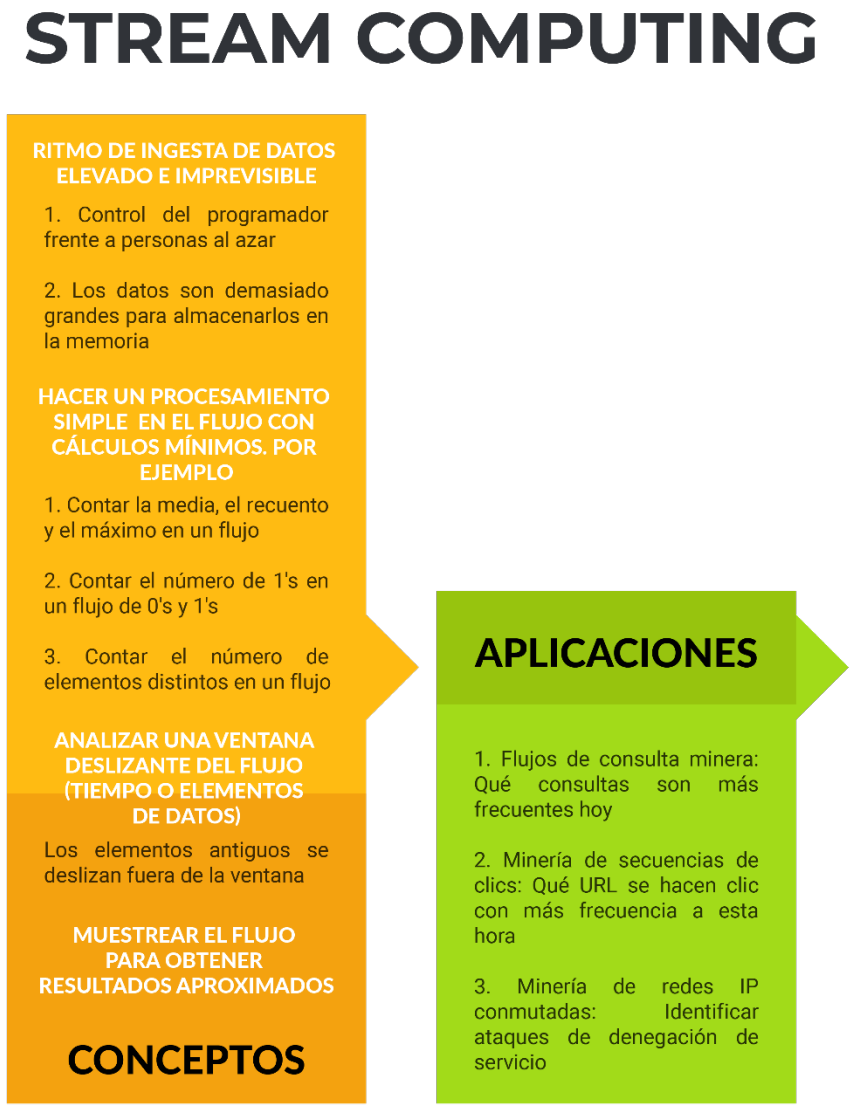


Figura 46. Conceptos y aplicaciones de Streaming.

Stream Computing

- Características del Algoritmo de Streaming
- Calcula las propiedades sin almacenar todo el flujo
- Limita el número de pasos aritméticos por cada nuevo elemento de datos
- N es tan grande que se querrá almacenar sólo O (log N)
- Filtrar el flujo para reducir el número de cosas que hay que hacer
- La eficiencia es esencial
- Consultas ad hoc y permanentes
- Filtro Bloom
- Es un algoritmo de Streaming muy popular
- Puede mostrar si una URL ha sido vista antes
- Utiliza una larga cadena de 1's y unas cuantas funciones hash
- Cada nueva entrada utiliza todas las funciones hash para establecer los bits en la cadena de 1's
- Si la nueva entrada tiene el hash de todos los 1's en la cadena, entonces esa entrada ha sido vista antes
- Crea algunos falsos positivos, pero ningún falso negativo

Stream Computing

- Apache Spark para Stream Computing

- Apache Spark es un motor integrado, rápido, en memoria y de propósito general para el procesamiento de datos a gran escala
- Spark es ideal para tareas de procesamiento iterativo e interactivo en grandes conjuntos y flujos de datos
- Spark consigue un rendimiento entre 10 y 100 veces superior al de Hadoop al funcionar con una construcción en memoria denominada "conjuntos de datos distribuidos resistentes" (RDD), que ayudan a evitar las latencias que conllevan las lecturas y escrituras en disco
- Spark permite a los programadores desarrollar cadenas de datos complejas y de varios pasos utilizando el patrón de gráficos acíclicos dirigidos (DAG)
- Spark está escrito principalmente en Scala, un lenguaje de alto nivel. Las bibliotecas integradas de Spark (para el aprendizaje automático, el procesamiento de gráficos, el procesamiento de flujos y el SQL) ofrecen un procesamiento de datos rápido y sin fisuras, además de una gran productividad para el programador
- Spark se ha convertido en una alternativa más eficiente y productiva para Hadoop. Es compatible con los sistemas de archivos y las herramientas de Hadoop

Stream Computing

Figura 47. Ejemplos de soluciones de stream computing.

Frameworks de Datos Masivos

BDPC™ Versión 092022

Frameworks de Apache

APACHE HADOOP

Funciona como un ecosistema de datos, es gratuito y de código abierto, usado para procesar grandes volúmenes de datos en bloques usando modelos de programación.

APACHE STORM

Procesa en tiempo real grandes cantidades de datos a través desde la creación de topologías llamadas macrodatos.

Su principal diferencia con otro framework orientado a mensajes como Kafka es que este es para procesamiento de mensajes en tiempo real y Kafka es un sistema distribuido de mensajería.

APACHE SPARK

Facilita el procesamiento de datos en paralelo a través de aprendizaje automático principalmente.

APACHE KAFKA

Sistema de mensajería. Este tiene varios componentes como son:

1. El ecosistema que es el compendio de aplicaciones que lo usan.

2. El clúster que es un sistema que se compone de diferentes corredores, temas y sus respectivas particiones.

3. Los consumidores de mensajes.

4. Los bróker que son el servidor Kafka que sirve como puente entre productores y consumidores.

5. Los tópicos que son la representación de un tipo similar de datos.

6. La partición de mensajes o datos.

7. El ZooKeeper que se utiliza para almacenar información sobre el clúster de Kafka, así como los detalles de los clientes consumidores. Existe un líder y sus seguidores, los cuales tienen comunicación el líder y cada ZooKeeper tiene su(s) bróker(s).

APACHE CASSANDRADB

Motor de almacenamiento orientado a columnas de alta escalabilidad y disponibilidad, se destaca por la replicación de datos en nodos. Por lo anterior, cuenta con un factor de replicación para hacer copias de datos en múltiples nodos para así tener garantía de la fiabilidad y tolerancia a fallos de lo que se almacena.

Se compone de un nodo, un data center, un clúster, un commit log (para recuperación en caso de alguna falla), una mem-table (una estructura de datos residente en memoria), el bloom filter ya explicado en el capítulo 11 y la SSTable que es un archivo de disco al que se descargan los datos de la Mem-Table cuando su contenido alcanza un valor de umbral.

Figura 48. Frameworks de Apache. Información adaptada de (JavaTPoint, s.f.).

Frameworks de Apache

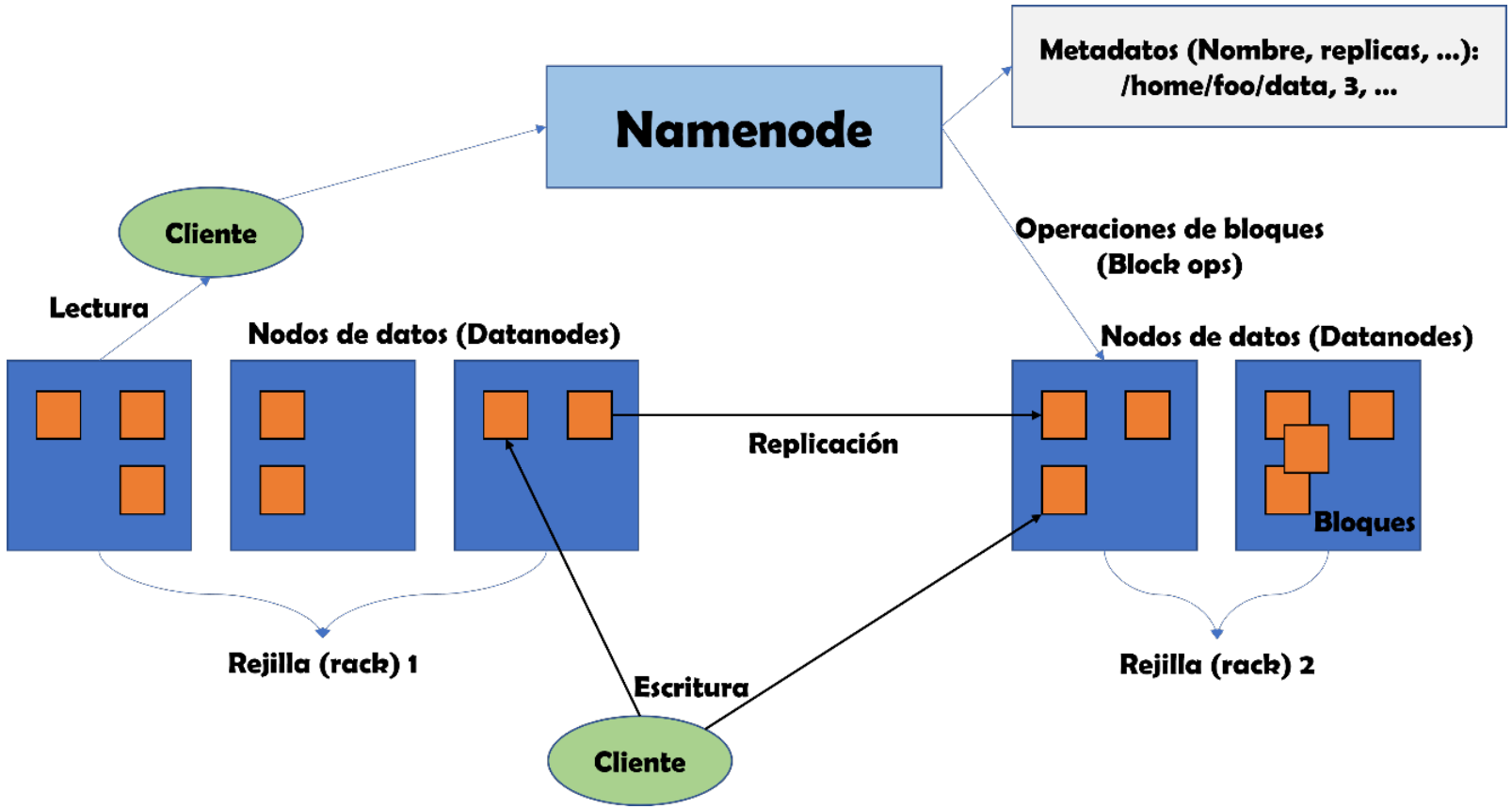


Figura 49. Arquitectura Apache Hadoop - HDFS. Información adaptada de (Haddop Apache, s.f.).

Frameworks de Apache

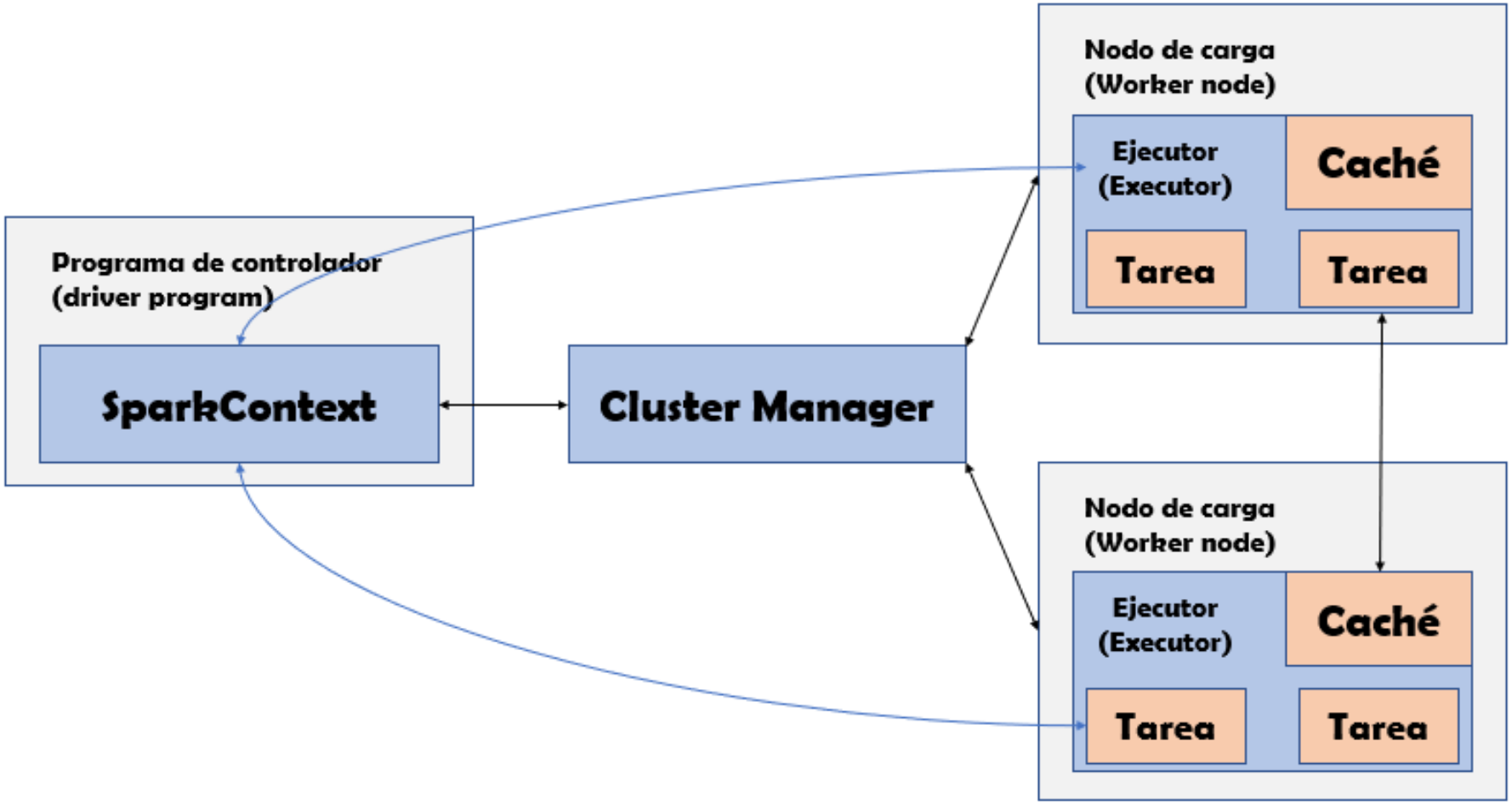


Figura 50. Arquitectura Apache Spark. Información adaptada de (Apache Spark, s.f.)

Frameworks de Apache

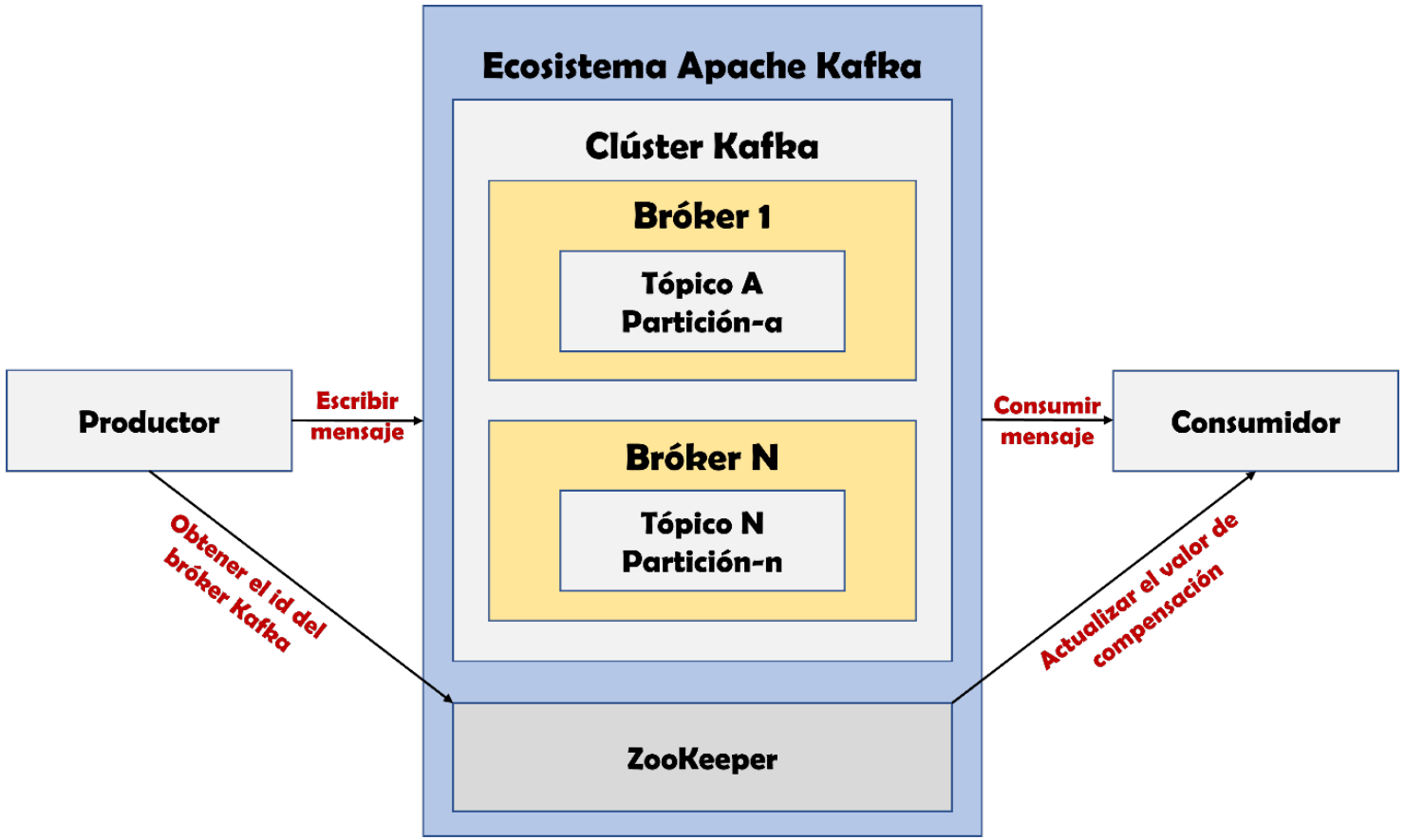


Figura 51. Arquitectura Apache Kafka. Información adaptada de (JavaTPoint, s.f.).

Frameworks de Apache

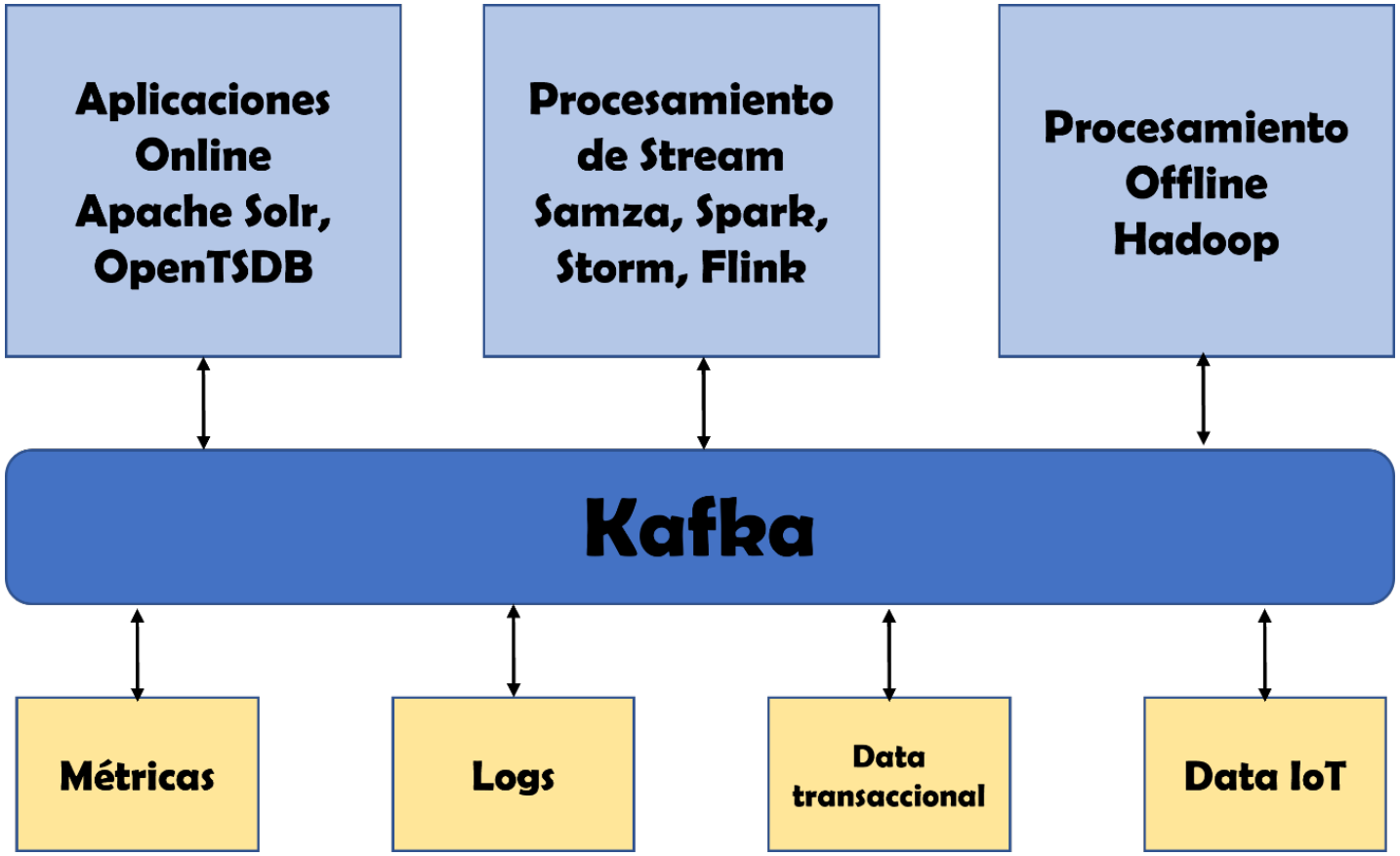


Figura 52. Ecosistema de Apache Kafka. Información adaptada de (JavaTPoint, s.f.).

Otros Frameworks de Datos Masivos

Infografía

Figura 53. Otros frameworks de datos masivos.

Computación en la Nube (Cloud Computing)

Cloud computing o computación en la nube es un modelo que permite una amplia flexibilidad y la utilización de una capacidad informática que se puede dividir en computación, almacenamiento, red, y que se puede adquirir sin necesidad de invertir en capacidad informática propia. Su capacidad y sus funciones dependerán del uso que el cliente desee darle de acuerdo con sus requerimientos y casos de uso. Para estos efectos, la computación en la nube cuenta con los siguientes beneficios:

- **Capacidad flexible:** La capacidad tiene una inmensa facilidad para ampliarse, o reducirse si así se requiere
- **Atractivo modelo de pago:** Su modelo de pago es comúnmente dado por uso, lo que implica que solo se consume lo que se ocupa y el tiempo de permanencia
- **Resistencia y seguridad:** Si un servidor individual y los recursos de almacenamiento fallan, no tiene una afectación directa al usuario porque hay un respaldo aislado para los clientes para que se maximice la seguridad de los datos

Computación en la Nube (Cloud Computing)

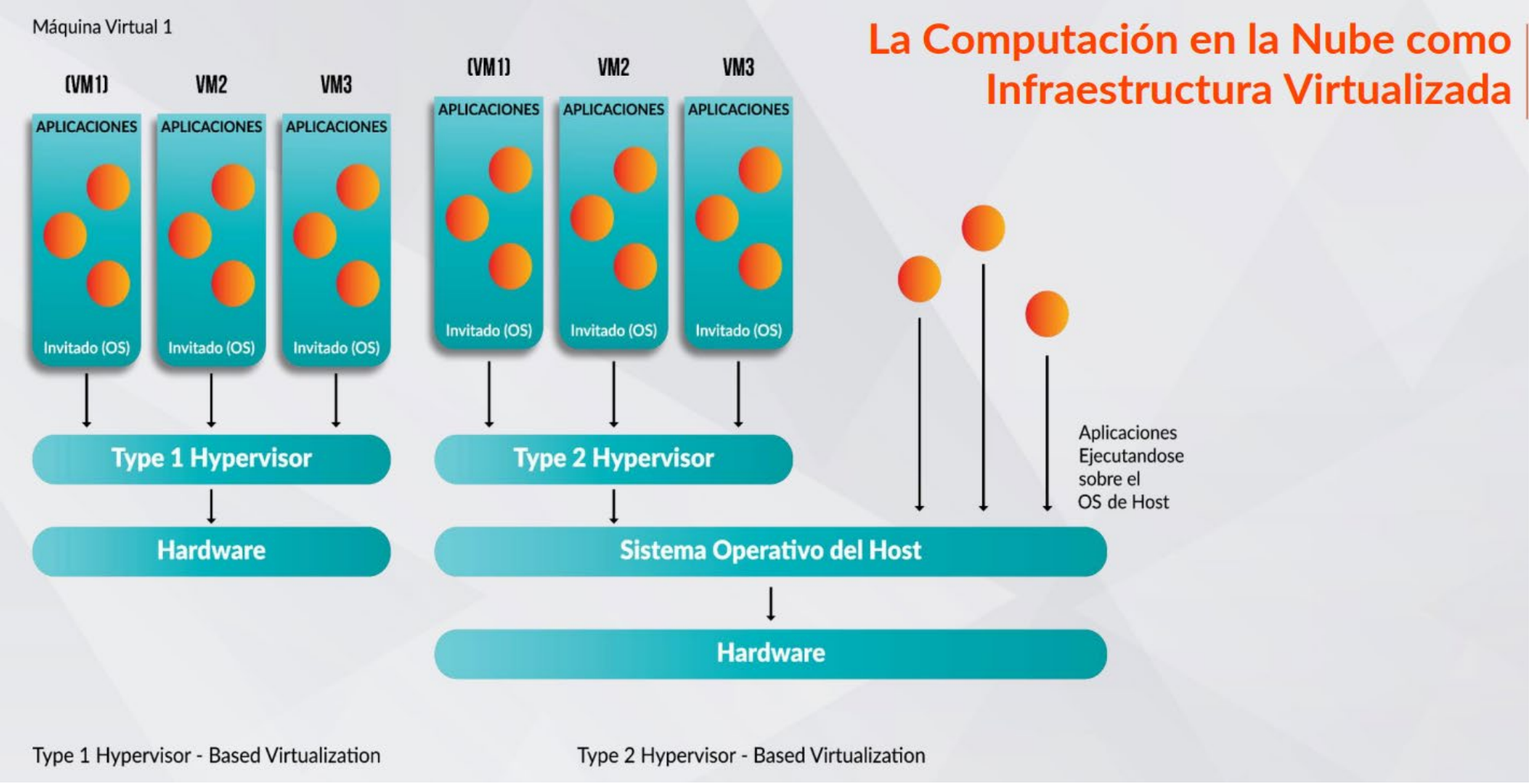


Figura 54. La computación en la nube como infraestructura virtualizada.

Computación en la Nube (Cloud Computing)

Tipos de servicio:

- **Nube pública:** Permite el acceso a servicios y recursos de computación cloud mediante internet.
- **Nube privada:** Para el acceso a servicios y recursos de computación en la nube con acceso restringido a usuarios de una organización
- **Nube híbrida:** Se combinan las características de servicios Cloud de tipo públicos y privados.

El Cloud para Big Data permite ofrecer servicios en la nube en tres modalidades según las necesidades de uso: 1) Infraestructura como servicio (IaaS), 2) Plataformas como servicio (PaaS), y 3) Software como servicio (SaaS).

Computación en la Nube (Cloud Computing)

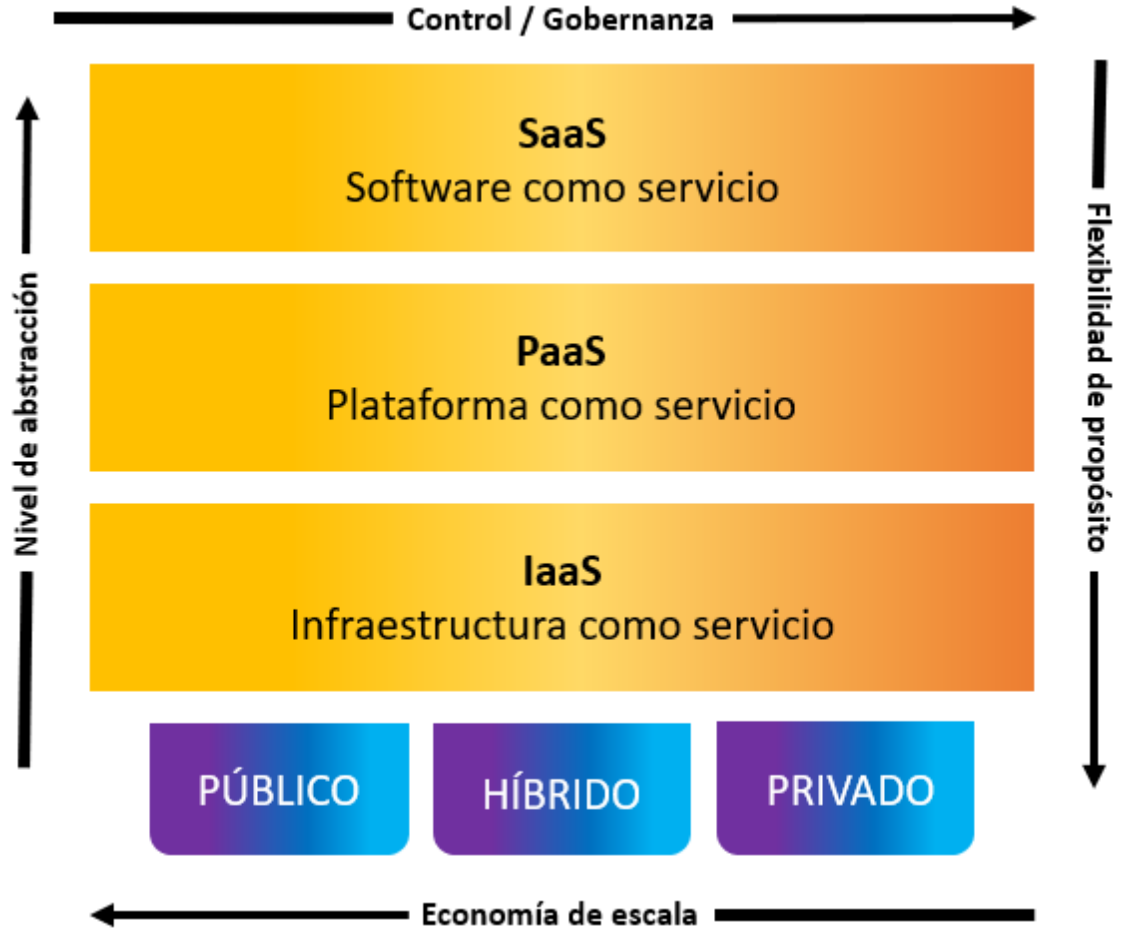


Figura 55. Modelo de computación en la nube por propiedad y rango de servicios.

Computación en la Nube (Cloud Computing)

Ejemplos de servicios de la nube en el mercado

- **Amazon EMR**
  - "Es la solución de macrodatos en la nube líder del sector destinada al procesamiento de datos a escala de petabytes, análisis interactivo y aprendizaje automático mediante el uso de marcos de código abierto, como Apache Spark, Apache Hive y Presto." (AWS, s.f.). Permite una creación de aplicaciones frameworks open-source cuyas opciones de ejecución están en clústeres personalizados de Amazon EC2, EKS, AWS Outposts o Amazon EMR Severless.

Computación en la Nube (Cloud Computing)

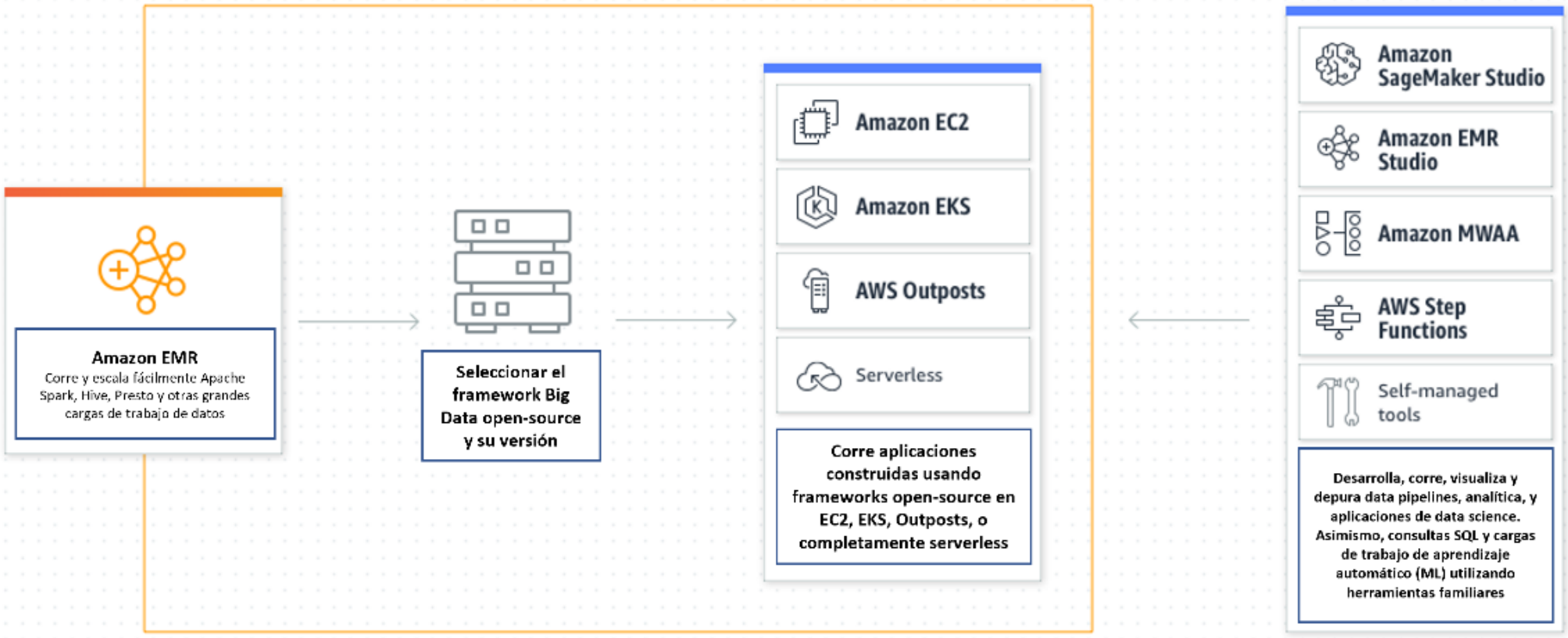


Figura 56. Funcionamiento de Amazon EMR. Información adaptada de (AWS, s.f.).

Computación en la Nube (Cloud Computing)

- **Oracle Big Data Service**
  - "Oracle Big Data Service es un servicio de Oracle Cloud Infrastructure diseñado para un conjunto diverso de cargas de trabajo y casos de uso de Big Data. Desde clústeres de corta duración utilizados para abordar tareas específicas hasta clústeres de larga duración que administran grandes lagos de datos, escalas de Big Data Service para cumplir con los requisitos de una organización a un bajo costo y con los más altos niveles de seguridad." (Oracle, s.f.).

Computación en la Nube (Cloud Computing)

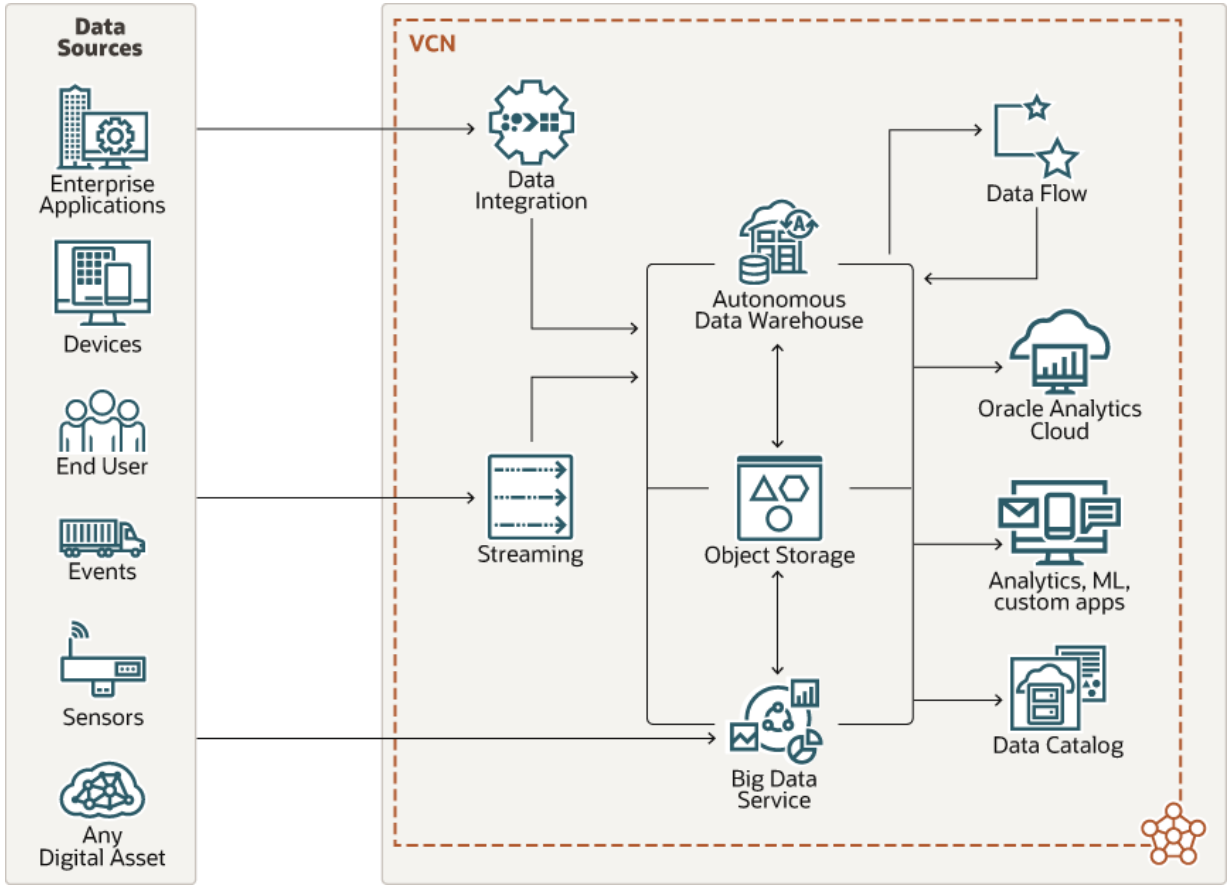


Figura 57. Arquitectura de la nube de Oracle. Obtenido de (Oracle, s.f.).

Computación en la Nube (Cloud Computing)

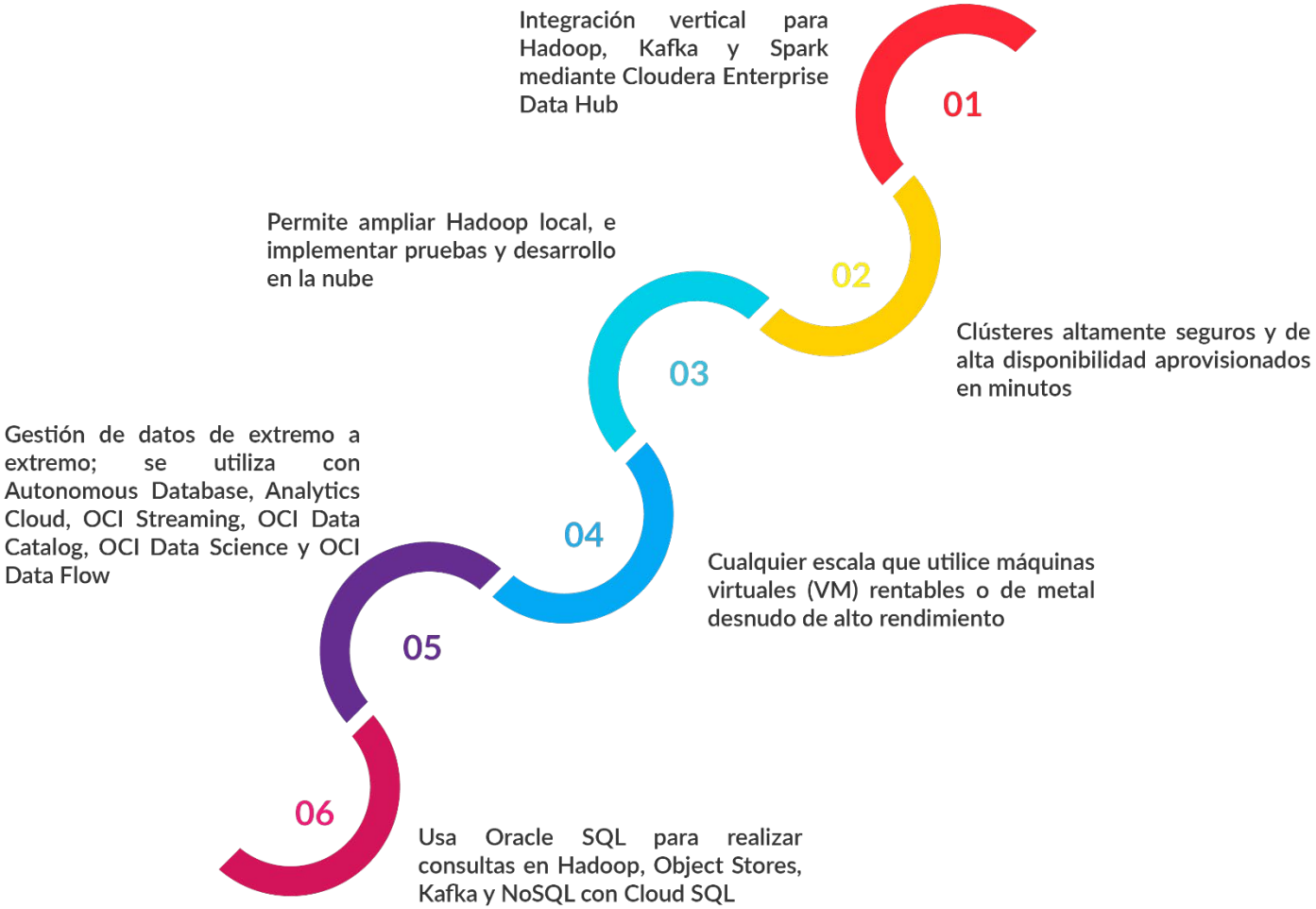


Figura 58. Características de Oracle Big Data Service. Información adaptada de (Oracle, s.f.).

Computación en la Nube (Cloud Computing)

- **Microsoft Azure**
  - Azure es el servicio de la nube de Microsoft que se divide en múltiples categorías, entre las que destacan la computación, móvil, almacenamiento, analítica, networking, integración, DevOps, seguridad, containers, IA, Blockchain y bases de datos. No obstante, entre otros, también destacan los servicios de Big Data que ofrece que son: Azure Synapse Analytics, HDInsight, Azure Databricks y Data Lake Analytics.

Computación en la Nube (Cloud Computing)

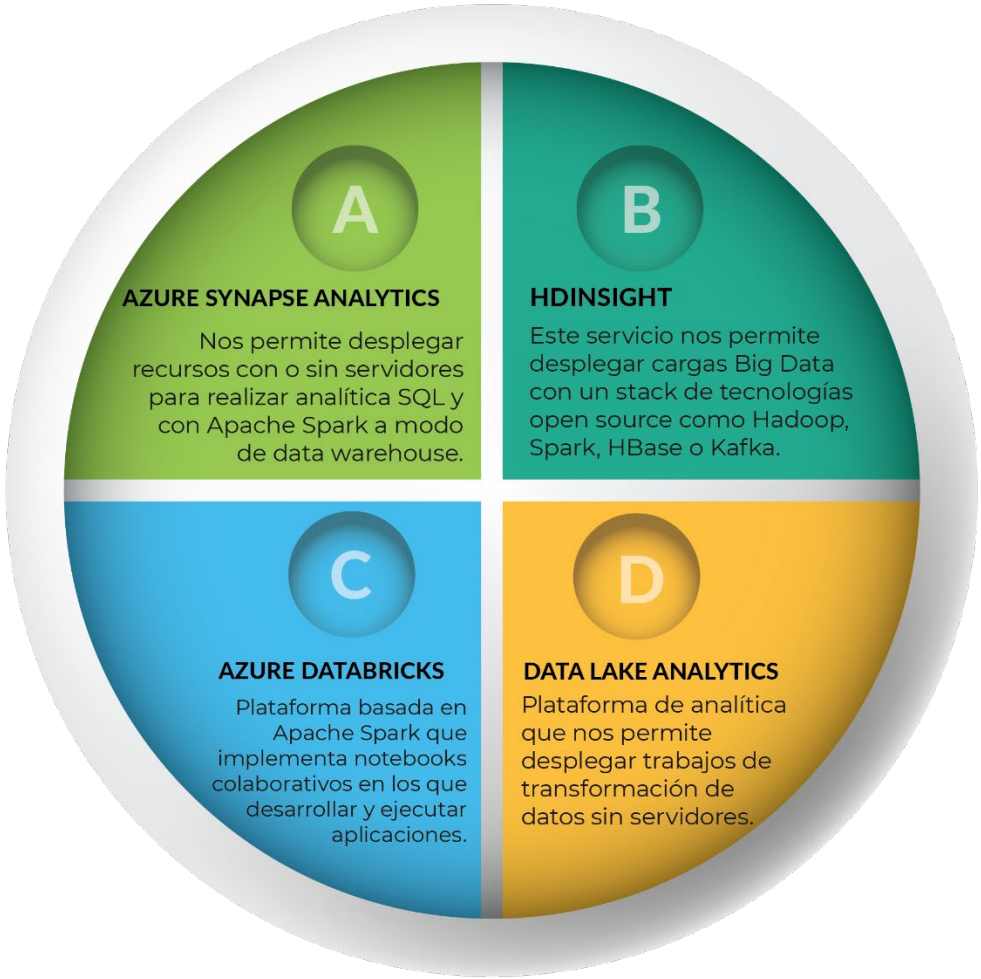


Figura 59. Servicios de Big Data de Microsoft Azure. Información adaptada de (AprenderBigData, 2022).

Computación en la Nube (Cloud Computing)

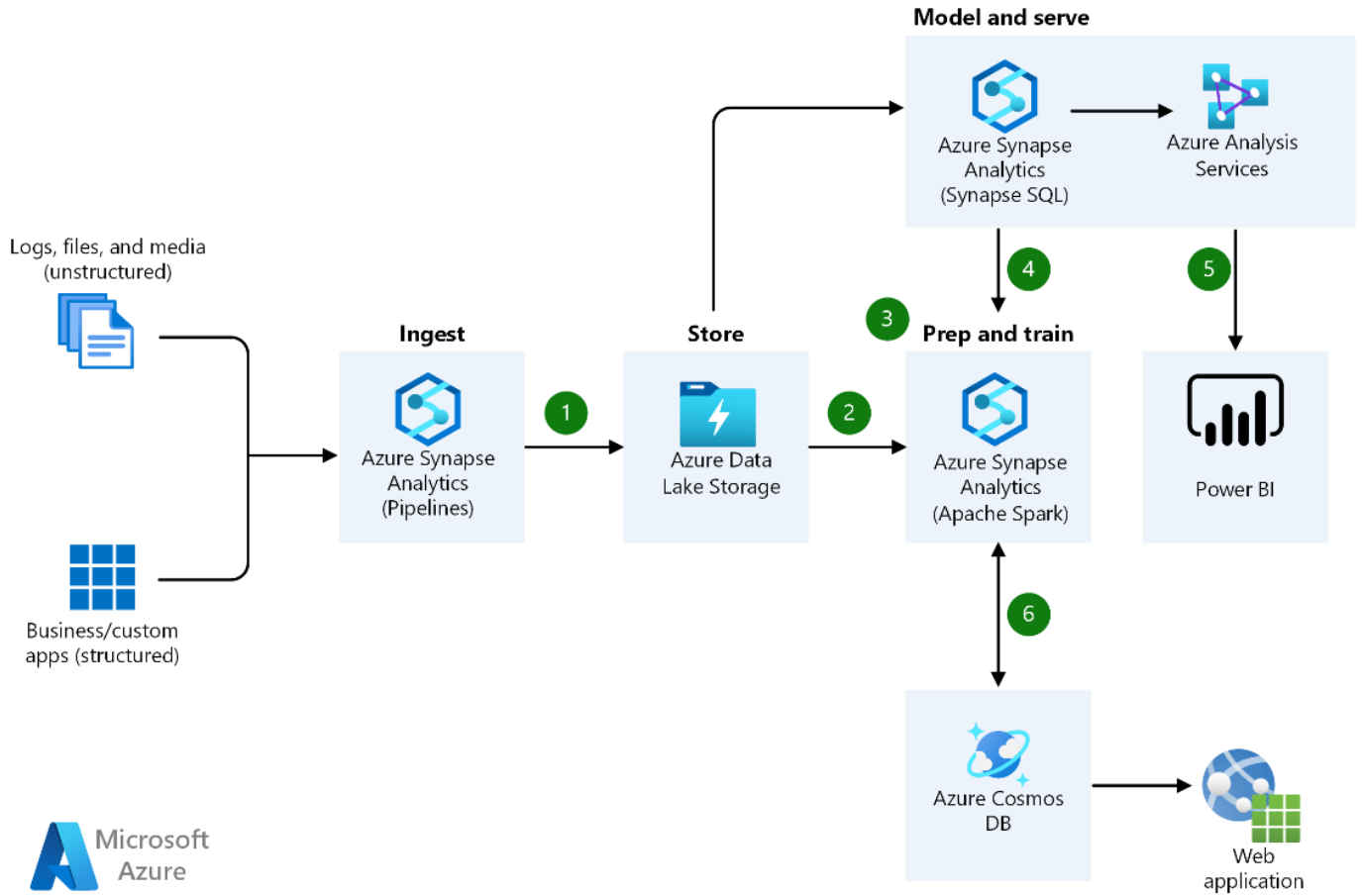


Figura 60. Arquitectura de análisis avanzado con Azure Synapse Analytics. Obtenido de (Microsoft, s.f.).

Computación en la Nube (Cloud Computing)

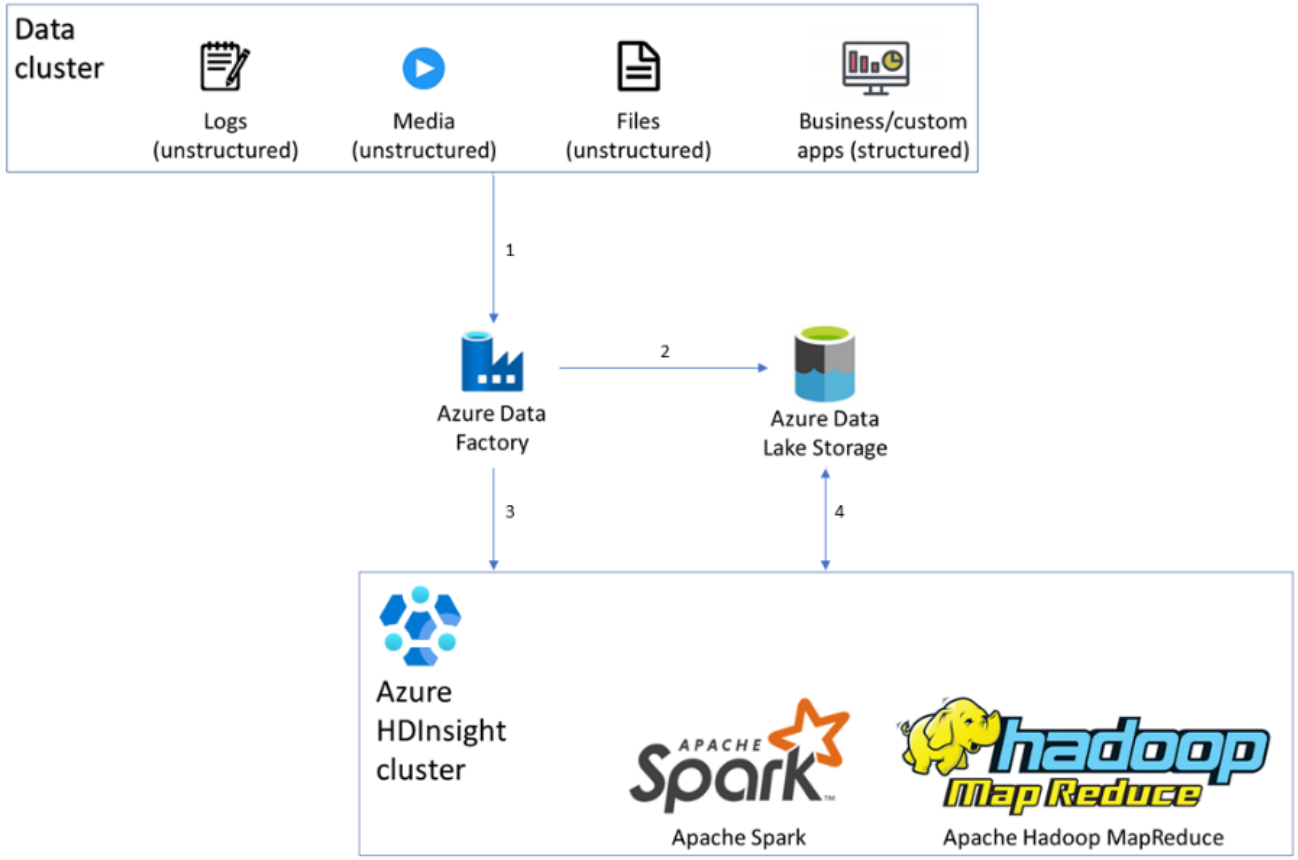


Figura 61. Arquitectura de extracción, transformación y carga de datos (ETL) mediante HDInsight. Obtenido de (Microsoft, s.f.).

Computación en la Nube (Cloud Computing)

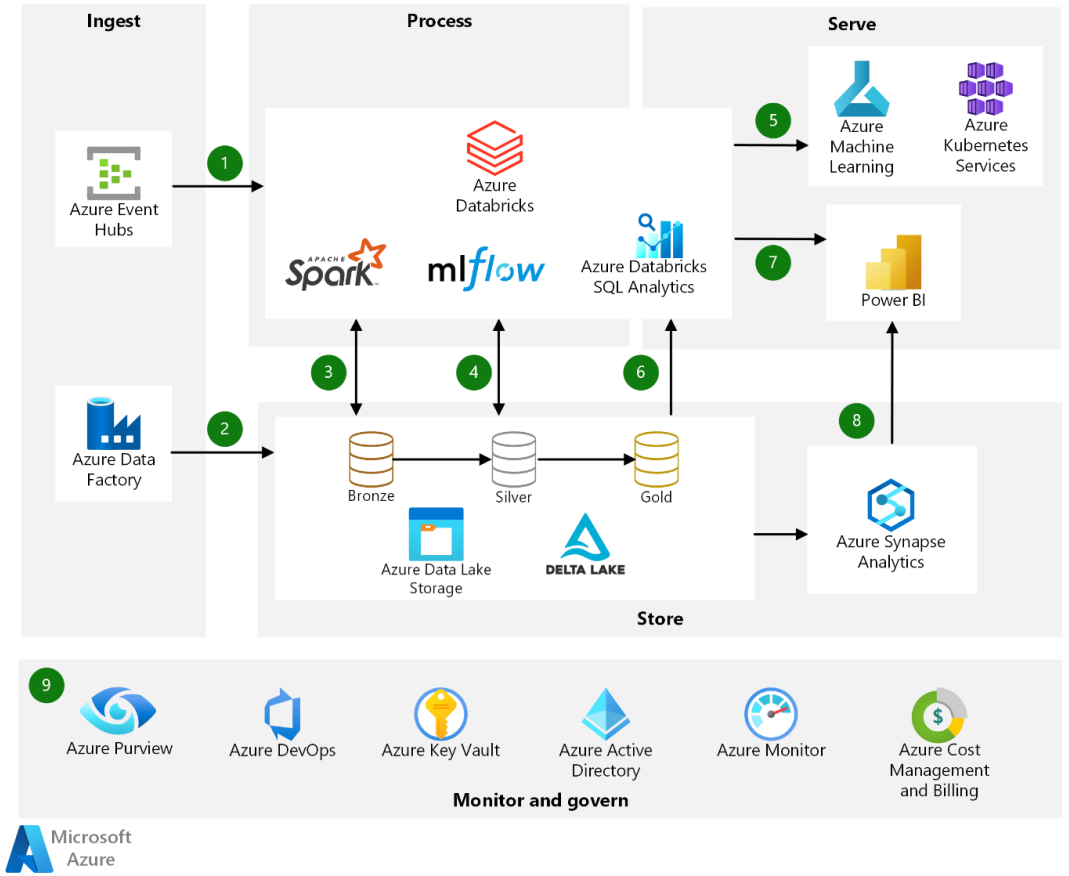


Figura 62. Arquitectura de análisis moderno con Azure Databricks. Obtenido de (Microsoft, s.f.).

Computación en la Nube (Cloud Computing)

- Google Cloud
  - El framework de la arquitectura de Google Cloud se divide en seis categorías o pilares, de los cuales el fundamental es el diseño de sistemas que es la base del framework que define el funcionamiento general de la arquitectura, mientras que los otros pilares son: Excelencia operativa; seguridad, privacidad y cumplimiento; confiabilidad; optimización de costos; y optimización de rendimiento (Google Cloud, s.f.).
  - Dentro de Google Cloud hay un servicio administrador por Spark y Hadoop denominado Dataproc, el cual puede aprovechar herramientas open-source para procesamientos por búsquedas, transmisiones, lotes y machine learning. Para todo lo anterior cuenta con un amplio abanico de integración con otros servicios de Google Cloud Platform como BigQuery, Cloud Storage, Cloud Bigtable, Cloud Logging y Cloud Monitoring y utiliza los clústeres tanto de Spark como Hadoop sin necesidad de asistencia de un software especializado o un administrador gracias a la consola de Google Cloud, el SDK de Cloud o vía API REST de Dataproc (Google Cloud, s.f.).

Computación en la Nube (Cloud Computing)

- BigQuery es un data warehouse que permite extraer analíticas de petabytes de datos (Google Support, s.f.)., además de permitir trabajar con lenguaje SQL y valerse de sus herramientas de machine learning integrado (BigQuery ML) que incluye su funcionalidad en la consola de Google Cloud, la herramienta de línea de comandos bq, la API REST de BigQuery y también en herramientas externas como lo son los notebooks Jupyter (Google Cloud, s.f.).
- Para una visualización de datos tienen un servicio llamado Google Data Studio que es una herramienta de BI gratuita de autoservicio que permite crear consultas y visualizaciones en BigQuery, así como generación de informes en Data Studio (Google Cloud, s.f.).

Computación en la Nube (Cloud Computing)

Figura 63. Comparación entre los servicios de nube de AWS, Azure, OCI y Google Cloud para la gestión de datos con Big Data. Información adaptada de (Oracle, s.f.).

Referencias

- Ashaari, M. A., Singh, K. S. D., Abbasi, G. A., Amran, A., & Liebana-Cabanillas, F. J. (2021). Big data analytics capability for improved performance of higher education institutions in the Era of IR 4.0: A multi-analytical SEM & ANN perspective. Technological Forecasting and Social Change, 173, 121119. doi:10.1016/j.techfore.2021.121119
- Azgomi, H., & Sohrabi, M. K. (2021). MR-MVPP: A map-reduce-based approach for creating MVPP in data warehouses for big data applications. Information Sciences, 570, 200–224. doi:10.1016/j.ins.2021.04.004
- Chaudhari, A.A. & Mulay, P. (2019). SCSI: Real-Time Data Analysis with Cassandra and Spark. In: Mittal, M., Balas, V., Goyal, L., & Kumar, R. (eds). Big Data Processing Using Spark in Cloud. Studies in Big Data, vol 43. Springer, Singapore. https://doi.org/10.1007/978-981-13-0550-4_11
- Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 165, 234–246. doi:10.1016/j.ijpe.2014.12.031

Referencias

- Gartner (s.f.). Gartner Glossary. https://www.gartner.com/en/information-technology/glossary/big-data
- Gudditti, V., & Venkata Krishna, P. (2021). Light weight encryption model for map reduce layer to preserve security in the big data and cloud. Materials Today: Proceedings. doi:10.1016/j.matpr.2021.01.190
- Kachris, C., & Tomkos, I. (2012). The rise of optical interconnects in data centre networks. 2012 14th International Conference on Transparent Optical Networks (ICTON). doi:10.1109/icton.2012.6253903
- Khan, M. A., Uddin, M. F., & Gupta, N. (2014). Seven V's of Big Data understanding Big Data to extract value. Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education. doi:10.1109/aseezone1.2014.6820689
- Li, X., Xue, F., Qin, L., Zhou, K., Chen, Z., Ge, Z., … Song, K. (2020). A recursively updated Map-Reduce based PCA for monitoring the time-varying fluorochemical engineering processes with big data. Chemometrics and Intelligent Laboratory Systems, 206, 104167. doi:10.1016/j.chemolab.2020.104167
- Russom, P. (2011). Big Data Analytics. In: TDWI Best Practices Report. https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

Referencias

- Cambridge Dictionary (s.f.). Data. https://dictionary.cambridge.org/es-LA/dictionary/english/data
- Cruz, N. P., Maña, M. J., & Mata, J. (s.f.). Aprendizaje Automático versus Expresiones Regulares en la Detección de la Negación y la Especulación en Biomedicina. http://www.uhu.es/noa.cruz/Cruz_Ma%C3%B1a_Mata.pdf
- Flanders, J., & Jannidis, F. (2015). Data Modeling. A New Companion to Digital Humanities, 229–237. doi:10.1002/9781118680605.ch16
- Gartner (s.f.). Semantic Data Model. https://www.gartner.com/en/information-technology/glossary/semantic-data-model
- GoodData (2022). What Is a Semantic Data Model? https://www.gooddata.com/blog/what-a-semantic-data-model/
- JavaTPoint (s.f.). Data Models. https://www.javatpoint.com/data-models
- JavaTPoint (s.f.). Data Processing in Data Mining. https://www.javatpoint.com/data-processing-in-data-mining
- Oracle (2022). 5 Understanding Data Sources - JD Edwards EnterpriseOne Tools Configurable Network Computing Implementation Guide. In: JD Edwards EnterpriseOne Tools Documentation Release 8.98 Update 4. https://docs.oracle.com/cd/E17984_01/doc.898/e14695/undrstnd_datasources.htm#gb92530a86fbac423_ef90c_10bd4799fff__7f4e
- Oracle (s.f.). ¿Qué es una base de datos? https://www.oracle.com/co/database/what-is-database/

Referencias

- Delua, J. (2021). Supervised vs. Unsupervised Learning: What's the Difference? https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning
- Rao, V. (2018). From data to knowledge. https://developer.ibm.com/articles/ba-data-becomes-knowledge-1/
- Singh, D. P. & Kaushik, B. (2022). Machine Learning Concepts and its applications for prediction of diseases based on drug behaviour: An extensive review. Chemometrics and Intelligent Laboratory Systems, 104637. https://doi.org/10.1016/j.chemolab.2022.104637
- AWS (2022). ¿Qué es una estrategia de datos? https://aws.amazon.com/es/what-is/data-strategy/#:~:text=A%20data%20strategy%20is%20a,amounts%20of%20raw%20data%20today
- Dynamic (2020). Historia del Big Data. https://www.dynamicgc.es/historia-del-big-data/
- Enterprise Big Data Framework (2019). Where does 'Big Data' come from? https://www.bigdataframework.org/short-history-of-big-data/
- IBM Cloud Education (2020). Artificial Intelligence (AI). https://www.ibm.com/cloud/learn/what-is-artificial-intelligence
- Phillips, A. (2021). A history and timeline of big data. https://www.techtarget.com/whatis/feature/A-history-and-timeline-of-big-data
- Rossi, B. (2016). How Industry 4.0 is changing human-technology interaction. https://www.information-age.com/industry-4-0-changing-human-technology-interaction-123463164/

Referencias

- Data Governance Institute (s.f.). Defining Data Governance. https://datagovernance.com/defining-data-governance/
- Google Cloud (s.f.) ¿Qué es el gobierno de datos? https://cloud.google.com/learn/what-is-data-governance?hl=es
- Olavsrud, T. (2021). Data governance: A best practices framework for managing data assets. https://www.cio.com/article/202183/what-is-data-governance-a-best-practices-framework-for-managing-data-assets.html
- Oracle (2020). Top big data analytics use cases. https://www.oracle.com/a/ocom/docs/top-22-use-cases-for-big-data.pdf
- Andoh-Baidoo, F. K., Chavarria, J. A., Jones, M. C., Wang, Y., & Takieddine, S. (2022). Examining the state of empirical business intelligence and analytics research: A poly-theoretic approach. In Information & Management (Vol. 59, Issue 6, p. 103677). Elsevier BV. https://doi.org/10.1016/j.im.2022.103677
- Castillo Romero, J. A. (2019). Big Data. IFCT128PO. IC Editorial.
- IBM (s.f.). ¿Qué es Business Intelligence? https://www.ibm.com/co-es/topics/business-intelligence
- Oracle (s.f.). ¿Qué es Big Data? https://www.oracle.com/co/big-data/what-is-big-data/#best-practices
- Oracle (s.f.). ¿Qué es un almacén de datos? https://www.oracle.com/co/database/what-is-a-data-warehouse/#:~:text=Los%20data%20warehouses%20est%C3%A1n%20dise%C3%B1ados,generar%20conocimientos%20basados%20en%20an%C3%A1lisis.
- Tableau (s.f.). Business intelligence: A complete overview. https://www.tableau.com/learn/articles/business-intelligence#what-is
- Yvanovich, R. (2018). Business Intelligence and Analytics - From A to Z (Part 3). https://blog.trginternational.com/business-intelligence-a-to-z-terms-glossary-part-3

Referencias

- Bolbolian Ghalibaf, M. (2020). Relationship Between Kendall's tau Correlation and Mutual Information. Revista Colombiana de Estadística, 43(1), 3–20. doi:10.15446/rce.v43n1.78054
- Hernández G., C. L. & Dueñas R., M. J. (2009). Hacia una metodología de gestión del conocimiento basada en minería de datos. CONGRESO INTERNACIONAL DE COMPUTACIÓN Y TELECOMUNICACIONES - COMTEL. http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-96.pdf?sequence=1
- Peralta, F. (2014). Proceso de Conceptualización del Entendimiento del Negocio para Proyectos de Explotación de Información. Revista Latinoamericana de Ingenieria de Software. 2. 273. 10.18294/relais.2014.273-306.
- Singular (s.f.). CRISP-DM: La metodología para poner orden en los proyectos. https://www.sngular.com/es/data-science-crisp-dm-metodologia/
- Ahmed, R., Shaheen, S., & Philbin, S. P. (2022). The role of big data analytics and decision-making in achieving project success. In Journal of Engineering and Technology Management (Vol. 65, p. 101697). Elsevier BV. https://doi.org/10.1016/j.jengtecman.2022.101697
- Barrionuevo, C. Ierache, J. & Sattolo, I. (2020).Reconocimiento de emociones a través de expresiones faciales con el empleo de aprendizaje supervisado aplicando regresión logística. 978-987-4417-90-9. http://sedici.unlp.edu.ar/bitstream/handle/10915/114089/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

Referencias

- Google Developers (2022). What is Clustering? https://developers.google.com/machine-learning/clustering/overview
- ISO (2019). ISO/IEC 20546:2019. https://www.iso.org/obp/ui/#iso:std:iso-iec:20546:ed-1:v1:en
- Li, L., Lin, J., Ouyang, Y., & Luo, X. (Robert). (2022). Evaluating the impact of big data analytics usage on the decision-making quality of organizations. In Technological Forecasting and Social Change (Vol. 175, p. 121355). Elsevier BV. https://doi.org/10.1016/j.techfore.2021.121355
- Lin, S., Lin, J., Han, F., & Luo, X. (Robert). (2022). How big data analytics enables the alliance relationship stability of contract farming in the age of digital transformation. In Information & Management (Vol. 59, Issue 6, p. 103680). Elsevier BV. https://doi.org/10.1016/j.im.2022.103680
- Chang, W., Boyd, D. & Levin, O. (2019). NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. Special Publication (NIST SP). National Institute of Standards and Technology, Gaithersburg, MD. [online]. https://doi.org/10.6028/NIST.SP.1500-6r2
- MongoDB (s.f.). Fundamentos de las bases de datos NoSQL. https://www.mongodb.com/es/nosql-explained
- Vega, J., Ortega, J. & Aguilar, L. (2015). Arquitectura Tecnológica Para Big Data. Revista Científica. 1. 7. 10.14483/udistrital.jour.RC.2015.21.a1.

Referencias

- Apache Spark (s.f.). Cluster Mode Overview. https://spark.apache.org/docs/latest/cluster-overview.html
- AprenderBigData (2022). Servicios de Big Data y Bases de Datos en Azure. https://aprenderbigdata.com/bases-de-datos-azure/
- AWS (s.f.). Amazon EMR. https://aws.amazon.com/es/emr/?c=a&sec=srv
- Google Cloud (s.f.). ¿Qué es BigQuery ML? https://cloud.google.com/bigquery-ml/docs/introduction?hl=es-419
- Google Cloud (s.f.). ¿Qué es Dataproc? https://cloud.google.com/dataproc/docs/concepts/overview
- Google Cloud (s.f.). Framework de la arquitectura de Google Cloud. https://cloud.google.com/architecture/framework?hl=es-419
- Google Cloud (s.f.). Visualiza datos con Data Studio. https://cloud.google.com/bigquery/docs/visualize-data-studio
- Google Support (s.f.). BigQuery. https://support.google.com/cloud/answer/9113366?hl=es
- Hadoop Apache (s.f.). HDFS Architecture. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
- JavaTPoint (s.f.). Apache Kafka Architecture. https://www.javatpoint.com/apache-kafka-architecture

Referencias

- JavaTPoint (s.f.). Apache Kafka Vs. Apache Storm. https://www.javatpoint.com/apache-kafka-vs-apache-storm
- JavaTPoint (s.f.). Cassandra Architecture. https://www.javatpoint.com/cassandra-architecture
- Microsoft (s.f.). Arquitectura de análisis avanzado. https://docs.microsoft.com/es-es/azure/architecture/solution-ideas/articles/advanced-analytics-on-big-data
- Microsoft (s.f.). Arquitectura de análisis moderno con Azure Databricks. https://docs.microsoft.com/es-es/azure/architecture/solution-ideas/articles/azure-databricks-modern-analytics-architecture
- Microsoft (s.f.). Extracción, transformación y carga de datos (ETL) mediante HDInsight. https://docs.microsoft.com/es-es/azure/architecture/solution-ideas/articles/extract-transform-and-load-using-hdinsight
- Oracle (s.f.). Compare OCI with AWS, Azure, and Google Cloud. https://www.oracle.com/cloud/service-comparison/
- Oracle (s.f.). Desarrollo de aplicaciones modernas - Big data y análisis. https://docs.oracle.com/es/solutions/big-data-and-analytics/index.html#GUID-04F64035-CDDA-4FBF-BCCB-A578032FA90B
- Oracle (s.f.). Getting Started with Oracle Big Data Service (HA) - Workshop Introduction and Overview. https://oracle.github.io/learning-library/data-management-library/big-data/bds/workshops/freetier/?lab=intro

COMPARTE Y VERIFICA TUS LOGROS DE APRENDIZAJE FÁCILMENTE

#BDPC #certiprof

certiprof®

¡Síguenos, ponte en contacto!

www.certiprof.com