UCAB UNIVERSIDAD CATÓLICA ANDRÉS BELLO

COMPARACIÓN ENTRE *BIG DATA* Y OTRAS ÁREAS. LA ARQUITECTURA, EL CICLO DEL DATO. ÁREAS DE APLICACIÓN, RETOS ACTUALES Y CASOS DE ÉXITO

OBJETIVOS

- Conocer la relación de *big data* con áreas de conocimiento que trabajan con datos, como son la inteligencia de negocio y el análisis de datos.
- Estudiar cómo una arquitectura de solución de *big data* puede verse como una posible evolución de la arquitectura de inteligencia de negocio y que incluye el análisis de datos.
- Aprender sobre el ciclo de los datos, los roles asociados a cada actividad y cuáles son las áreas de aplicación de *big data*.

TABLA DE CONTENIDO

INTRODUCCIÓN

Desde el surgimiento de *big data* su implementación en las organizaciones ha generado elementos de discusión con respecto a áreas que se han desarrollado de manera paralela, como son la inteligencia de negocio y el análisis de datos. Es necesario tener una perspectiva integradora que permita llevar adelante un proyecto capaz de unir las diferentes visiones necesarias para la empresa, lo cual se puede ver a través de la identificación de los cambios en la arquitectura de solución.

Así mismo, es importante determinar y conocer la cadena de valor de *big data* y el ciclo que el dato estaría recorriendo, las características del almacenamiento en *big data* (llamado "lago de datos") para, finalmente, discutir sobre las técnicas de procesamiento del flujo de datos y cómo estos pueden impactar en una arquitectura de solución de *big data.*

01 Relación de *Big data* con Inteligencia de Negocio y la Analítica de Negocio

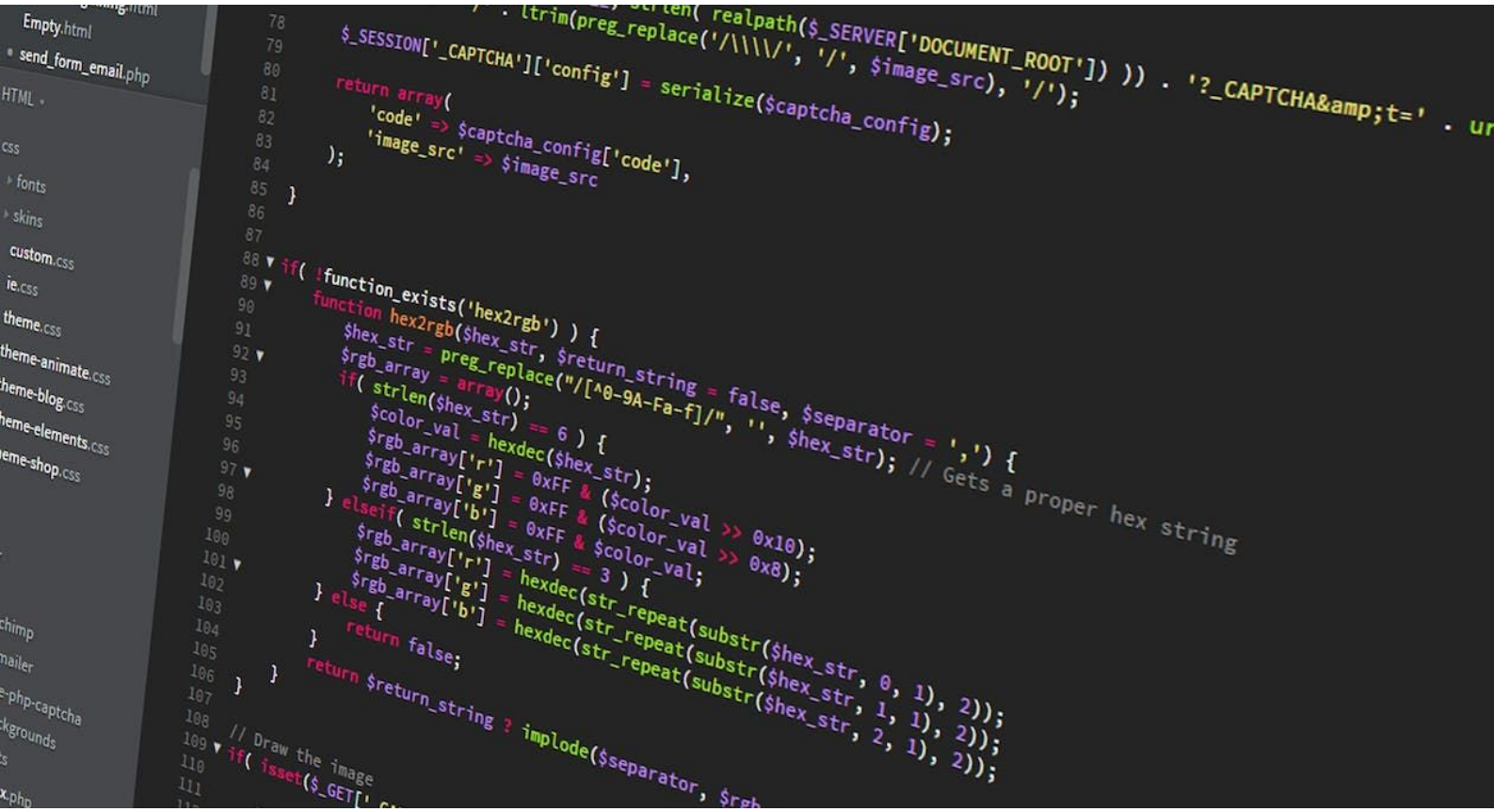
Podemos comenzar afirmando que *big data*, inteligencia de negocio (BI) y analítica de negocio (BA) son tres áreas en el ámbito organizacional que comparten un **objetivo común**: permitir el análisis de datos, con el fin de extraer la mayor información posible.

El término *Business Intelligence*, (inteligencia de negocio en español) fue acuñado en 1958 por el investigador de IBM, Hans Peter Luhn, en el artículo "A Business Intelligence System", donde lo definía como: "la habilidad de aprender las relaciones de hechos presentados de forma que guíen las acciones hacia una meta deseada" (Luhn, 1958, p. 314).

La inteligencia de negocios se vio potenciada en el año 1962, con la aparición del concepto de OLAP (procesamiento analítico en línea), acuñado por el canadiense Kenneth Iverson, que supuso un importante avance en la analítica de datos. Otro hito importante en la administración de datos fue la creación del concepto de bases de datos en 1969, el cual se asentó en la década de los setenta, y el desarrollo teórico y práctico de tan importante disciplina. En los años 80 apareció el concepto de *Data Warehouse* (almacenes de datos), junto con las metodologías de Bill Inmon y Raph Kimball, quienes se denominan las principales autoridades en esta área.

Fue en 1989 cuando Howard Dresden, investigador de la consultora Gartner, hizo una de las primeras definiciones de inteligencia de negocios: "conceptos y métodos para mejorar las decisiones de negocio mediante el uso de sistemas de soporte basados en hechos".

Por otro lado, **la analítica de negocios** es un proceso asistido por tecnologías mediante el cual el *software* analiza los datos para predecir lo que sucederá (análisis predictivo) o lo que podría suceder adoptando un cierto enfoque (analítica prescriptiva). El análisis de datos se complementa con otros dos tipos de análisis: descriptivo y de diagnóstico, ambos asociados directamente a la inteligencia de negocios tradicional.



En los últimos años, se han desplegado las metodologías y tecnologías de *Big Data*, por el crecimiento exponencial de datos presentes en las organizaciones y empresas. La era de los grandes volúmenes de datos (*Big Data*), su tratamiento, su explotación y la conversión de datos en conocimiento para una toma de decisiones efectiva. Las empresas necesitan obtener valor del gran volumen y variedad de información. Así han aparecido las nuevas tendencias de analítica de *Big Data*, como un proceso de examen de los grandes volúmenes de datos para descubrir patrones ocultos, correlaciones desconocidas y otra información de interés que se pueden utilizar para tomar mejores decisiones.

Estos pueden parecer conceptos similares ya que ambos comparten el mismo principio: aprovechar de la mejor manera la información para poder tomar mejores decisiones.

En la prensa generalista y en la prensa económica o tecnológica especializada, se suelen utilizar los tres términos, bien de modo diferenciado o bien como sinónimos. La realidad es que los tres conceptos conviven en consultoras, medios de comunicaciones, proveedores de software, desarrolladores de aplicaciones, etc. Es difícil encontrar semejanzas y diferencias, pero trataremos de hacerlo en este caso desde el punto de vista de que las tres disciplinas sirven para dar soporte a la toma de decisiones A pesar de ello, tienen sutiles diferencias en cuanto a 4 conceptos claves:

- qué datos analizan,
- dónde se almacenan,
- qué hacen con la información y
- qué variable estudia cada uno.

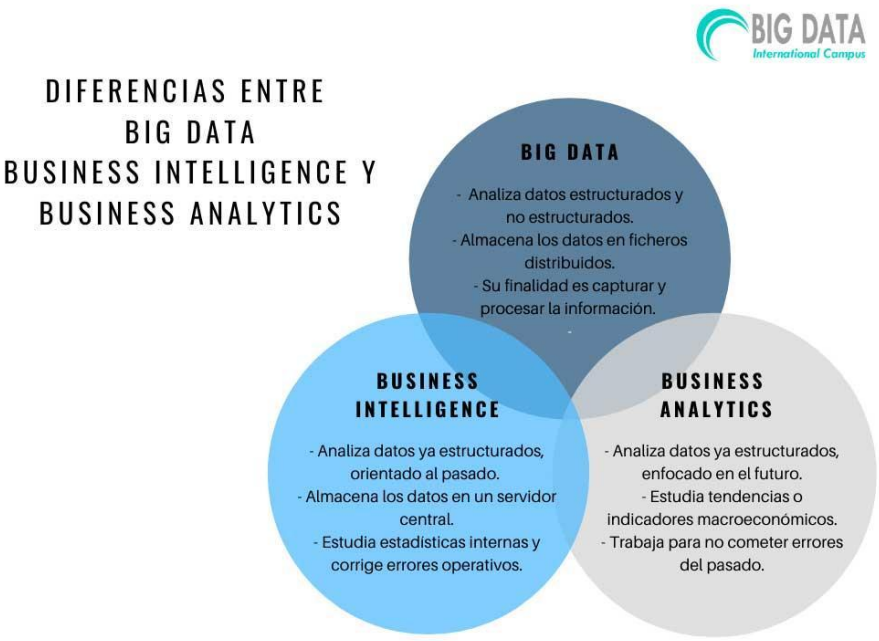
Por un lado, **la inteligencia de negocio** está orientada al pasado (que se representa con el espejo retrovisor). A través de informes mensuales o semanales se examina el recorrido histórico de la empresa y así se comprende su desarrollo. Se responden preguntas tales como: ¿Cómo va el negocio? ¿Se han cumplido las metas? ¿Cuál es su estado de salud operativa y financiera? Para ello se usan los indicadores que se agrupan en cuadros de mando, tal como el tablero de control de un automóvil que nos indica el nivel de gasolina, la temperatura, si hay una puerta abierta o una luz prendida, etc. Con esto podemos conocer el estado actual de nuestro auto y, además, nos permite tomar decisiones como reducir o aumentar la velocidad, parar para comprar gasolina o refrigerante, entre otros.

Con los instrumentos utilizados en BI se accede a conjuntos de datos preparados, debidamente clasificados y almacenados, que provienen en su mayoría de las bases de datos operativas del negocio como son los ERP, CRM, entre otros. Gracias a esto se examina la información y se encuentran patrones analíticos. Por consiguiente, se puede decir que la inteligencia de negocio únicamente analiza datos estructurados y los almacena en un servidor central, los cuales se proceden a estudiar para llegar a conclusiones que ayudan a tomar decisiones.

Por otro lado, **la analítica de negocio** (BA, *Business Analytic*) se enfoca en el futuro, es decir, facilita la creación de una visión futurible, basada en modelos predictivos o prescriptivos que influyen en la elección de nuevos caminos y estrategias. El BA no estudia estadísticas internas como el BI, sino que se sirve de diferentes fuentes: tendencias o indicadores macroeconómicos. Otra importante distinción con BI, no menos importante, está relacionada con el uso de la información que hace cada una. Ambas buscan optimizar los procesos: la inteligencia de negocio busca corregir errores operativos y la analítica de negocio trabaja con el objetivo de no cometer esos fallos en el futuro.

Por último, se entiende por *big data* al grupo de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos, que frecuentemente, pero no siempre, viene definida por volumen, velocidad y variedad (Díaz, 2011).

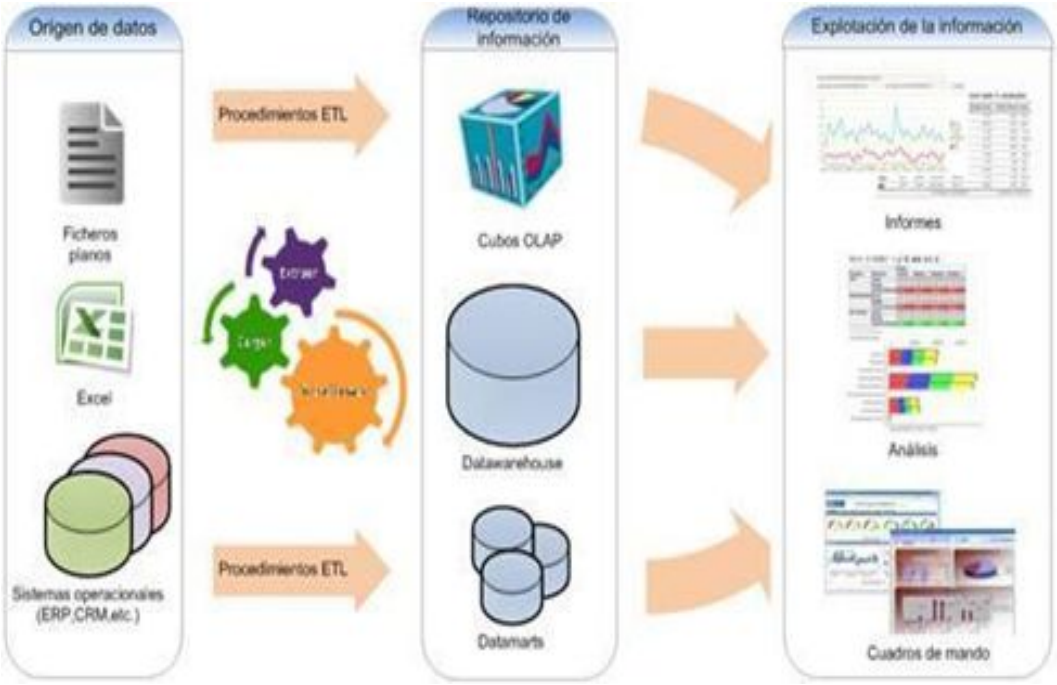
En cuanto a qué tipo de datos analiza, *big data* trabaja con una **gran cantidad de datos** provenientes de varias fuentes, siendo estos estructurados, semiestructurados y no estructurados; en contraposición, la inteligencia de negocio solamente realiza análisis de datos estructurados. *Big data*, a diferencia de la inteligencia de negocios, puede procesar datos tanto en tiempo real como en formato offline y almacenarlos en sistemas distribuidos como hadoop, el cual se estudiará en este curso.



Fuente: Diferencias entre *big data*, *business inteligence* y *business analytics.* Extraída de *Big Data* International Campus (2018)

Para comprender mejor las similitudes y diferencias, compararemos la arquitectura de BI y la de *big data*: la primera ha sido definida como un marco de trabajo (*framework*) que detalla los diferentes componentes del sistema de inteligencia de negocios, tales como datos, personas, procesos, tecnologías y gestión/administración, y la forma en que estos componentes se han de combinar y coordinar para asegurar el correcto funcionamiento del sistema.

La información contenida en una arquitectura de BI es el conjunto de tipos de datos que necesitan ser recolectados, los métodos que se utilizan para analizar los datos y el modo en que se presenta la información necesaria.



Fuente: Business Intelligence. Extraída de InnoWiki (2014)

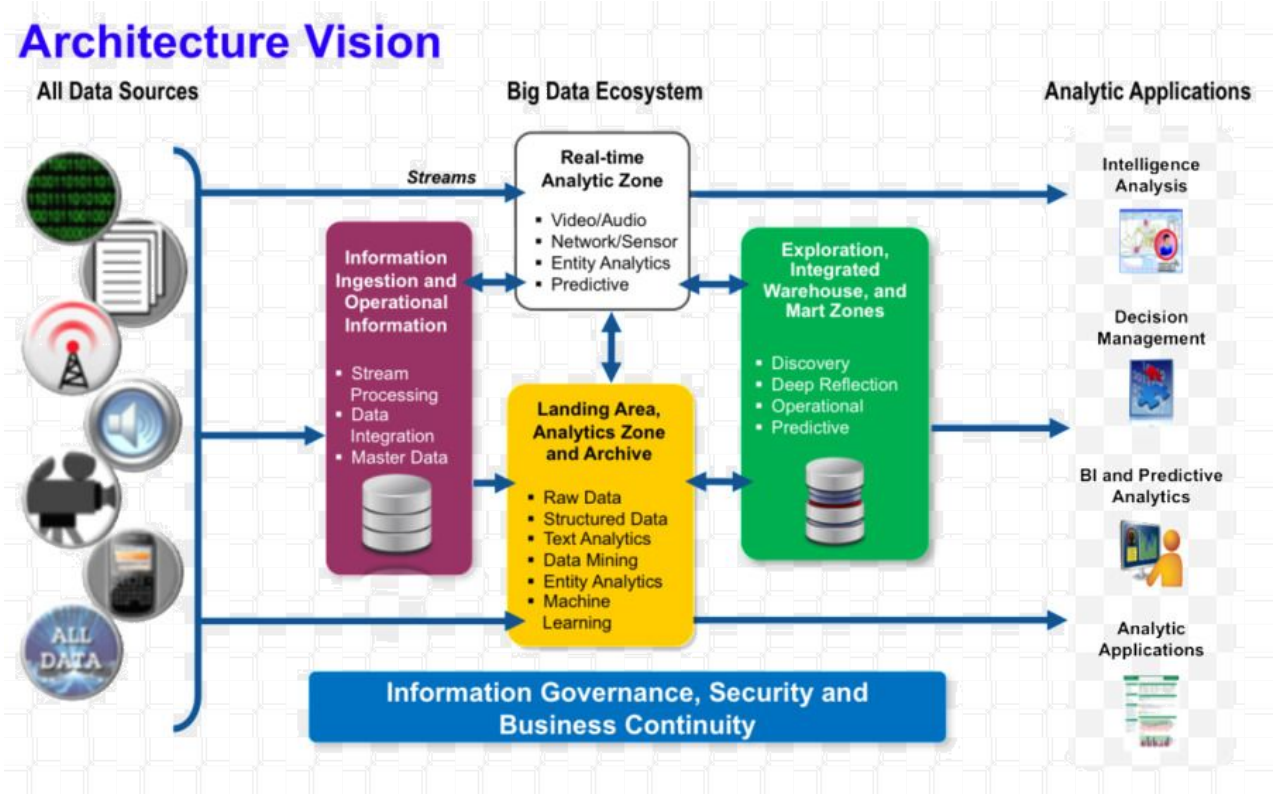
En la arquitectura BI se distinguen los siguientes componentes:

- **Origen o fuentes de datos:** las distintas fuentes de información origen sobre la que se empieza a montar el sistema BI. Estas pueden ser múltiples con distintos formatos.

- **ETL:** sobre las fuentes de información origen se montan los procesos ETL (extracción, transformación y carga), los cuales recogen la información de las fuentes de datos origen, realizan las transformaciones oportunas y cargan la información en un nuevo repositorio de información (*Datawarehouse*), adaptado para poder realizar sobre él la exploración de la información.

- **Exploración de la información:** conjunto de herramientas que permiten recuperar la información del *Datawarehouse* adaptada a las necesidades que se requieran. Un sistema BI va a mostrar la información de tres formas diferentes, dependiendo del tipo de usuario y nivel de información: informes, análisis y cuadros de mando.

Compara la arquitectura de BI con una arquitectura de *big data* como la que puedes ver a continuación:

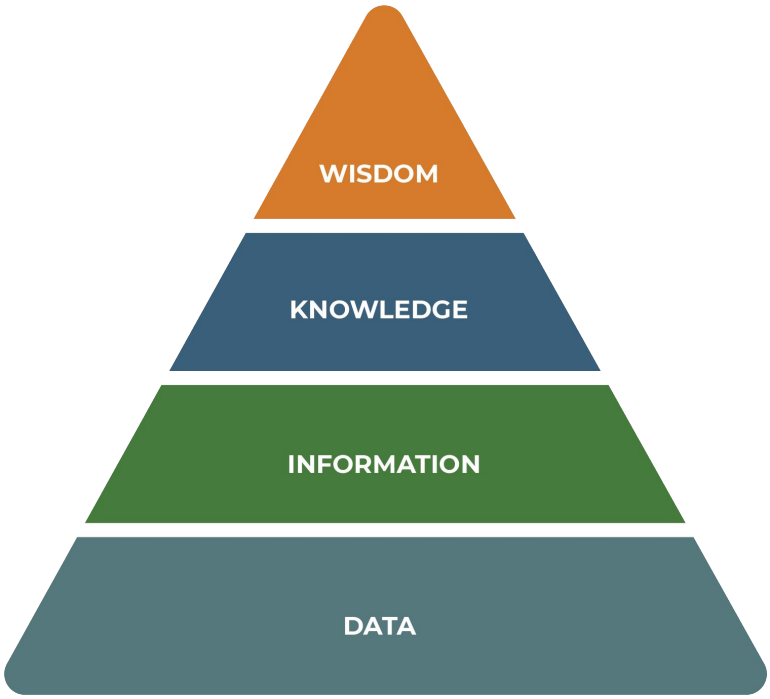


Fuente: Arquitectura de datos. Extraída de PNGWING (s.f.)

En esta arquitectura **los datos se extraen** de las fuentes de datos, que pueden ser internas o externas a la organización, y se trasladan a una zona de ingestión e información operacional o a una zona de análisis de tiempo real donde, por un lado, se almacenan en un lago de datos y por otro, se procesan para estructurarse en los *DataWarehouses* y *DataMarts*. Esta luego es analizada usando herramientas de BI o BA para luego llevarse a instrumentos de visualización que faciliten la comprensión de los resultados y permitan la toma de decisiones, todo bajo una política de gobernanza y seguridad de datos.

Por lo tanto, se puede observar que *big data* agrega elementos, capacidad y funcionalidad a lo que sería una solución típica de inteligencia de negocio, a tal nivel que hay expertos como Joyanes Aguilar (2019) que consideran que "*Big Data* es la evolución de Inteligencia de Negocios" (p. 35).

Dentro del campo de la **gestión empresarial**, las cadenas de valor se han utilizado como una herramienta de apoyo a la toma de decisiones, para representar la cadena de actividades que realiza una organización con el fin de entregar un producto o servicio de valor para el mercado (Porter 1985). La cadena de valor categoriza las actividades genéricas que agregan valor en una organización, lo que permite comprenderlas y optimizarlas. Una cadena de valor se compone de una serie de subsistemas, cada uno con entradas, procesos de transformación y salidas.



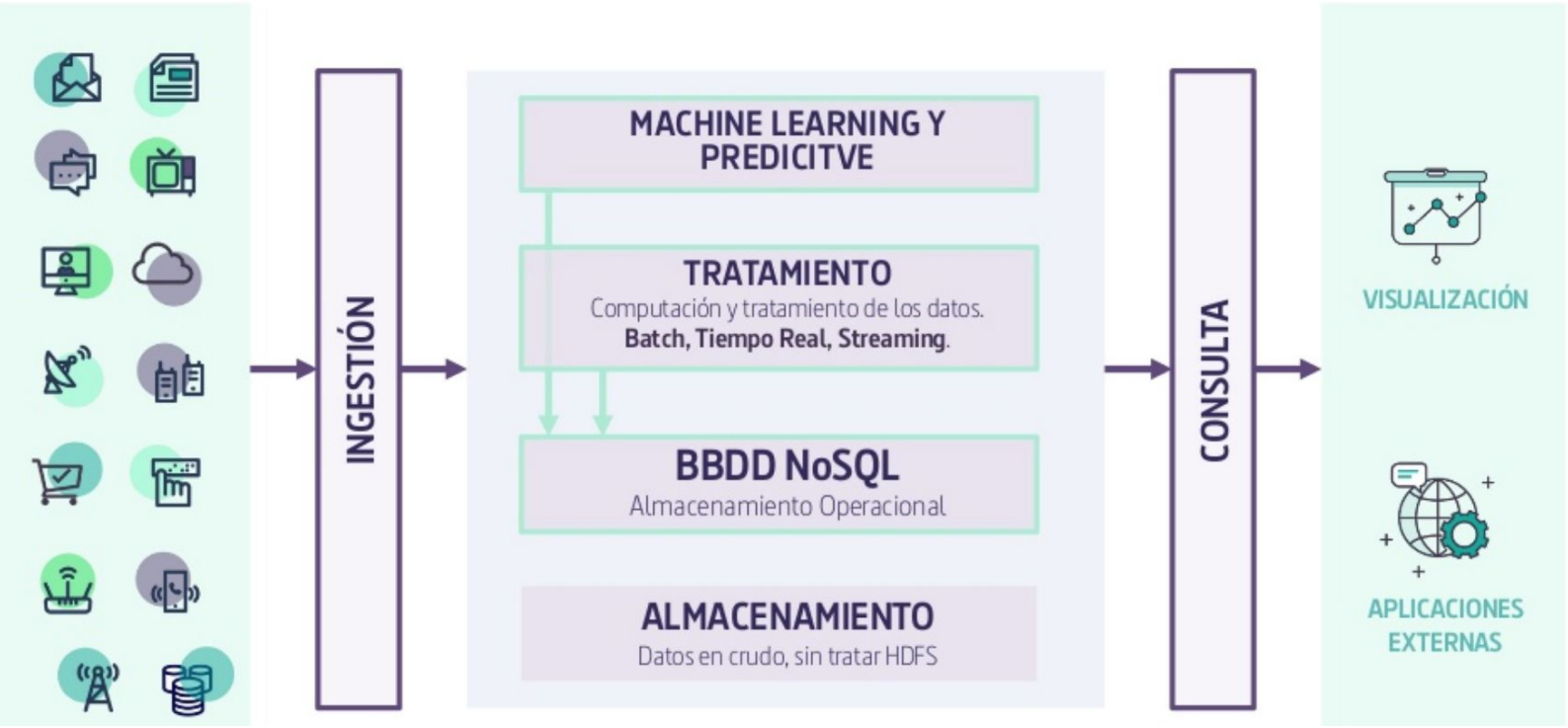
Fuente: The Data-Information-Knowledge-Wisdom hierarchy pyramid. Extraída de Jony et al. (2016)

En 1988, R. L. Ackoff especificó por primera vez la cadena de valor de los datos, que era una jerarquía basada en la filtrado, la reducción y la transformación, que mostraba cómo los datos conducen a la información, al conocimiento y finalmente a la sabiduría.

La cadena de valor de datos es un modelo que considera los datos como una materia prima y como un recurso importante en el negocio. En esta, el flujo de información se describe como una serie de pasos necesarios para generar valor y conocimientos útiles a partir de los datos.

En la cadena de valor que se ilustra anteriormente, se puede utilizar para modelar las actividades de alto nivel que componen un sistema de información. La cadena de valor de *Big Data* identifica las siguientes actividades clave de alto nivel:

Fuente elaboración propia basado en (Aguilar, 2019) (Faroukhi, Alaoui, Gahi, & Amine, 2020)



Previamente a la recolección de datos, se requiere de una etapa de identificación de las fuentes de datos, que es muy importante en la decisión de la arquitectura, ya que implica identificar las diferentes fuentes de datos y su clasificación en función de su naturaleza y tipos.

Los aspectos que se han de considerar en la **identificación de las fuentes de datos** son:

- Identificar las fuentes internas y externas.
- Calcular la cantidad de datos detectada (por ingerir) de cada fuente de datos.
- Identificar los mecanismos de obtención de datos (push o pull).
- Determinar el tipo de fuente de datos (generadas por máquinas o por personas, archivos, bases de datos de la empresa, datos web).
- Determinar el tipo de datos: estructurado, no estructurado o semiestructurado.

En esta etapa la disponibilidad, cantidad y accesibilidad definen el valor de las fuentes de datos, lo que significa que, si los datos de las fuentes son fácilmente accesibles, tienen un valor más alto.

La adquisición (ingesta) de datos

Se refiere a la forma en que se pueden **recibir y recopilar los datos** y se ha convertido en una etapa de gran interés en el proceso de *Big Data*, ya que existen numerosos datos públicos que se producen en enormes cantidades, numerosos dispositivos desperdigados por todo el planeta que emiten, procesan y recogen información de las más diversas actividades (posicionamiento de individuos y vehículos, niveles de contaminación, temperaturas,); de igual forma, infinidad de dispositivos móviles que también emiten y capturan datos, etcétera.

También representa un desafío en términos de requisitos de infraestructura. La infraestructura necesaria para respaldar la adquisición de *Big data* debe ofrecer una latencia baja y predecible; ser capaz de manejar volúmenes de transacciones muy altos en un ambiente distribuido y soportar estructuras de datos flexibles y dinámicas.

Esta fase consiste en identificar el modo de flujo de datos y como serán tratados al conectarse a plataformas de generación. Este flujo de datos podría ser:

- **El modo de carga por lotes** se puede realizar en grandes conjuntos de datos, agrupados en un intervalo de tiempo definido. A menudo se usa para fuentes de datos de sistemas heredados con procesos de prueba o cuando los flujos de datos no se pueden entregar técnicamente.

- **El modo de carga de flujo en entradas de datos continuas**. Debe funcionar en tiempo real o casi en tiempo real y tiene una tasa de carga más rápida que la tasa de datos entrantes.

- **El modo de carga de micro lotes** permite dividir los flujos de entrada en micro lotes. Como resultado, los datos se obtienen casi en tiempo real.

Estos se almacenan en forma de datos crudos en lo que se denomina un lago de datos.

Preprocesamiento de datos

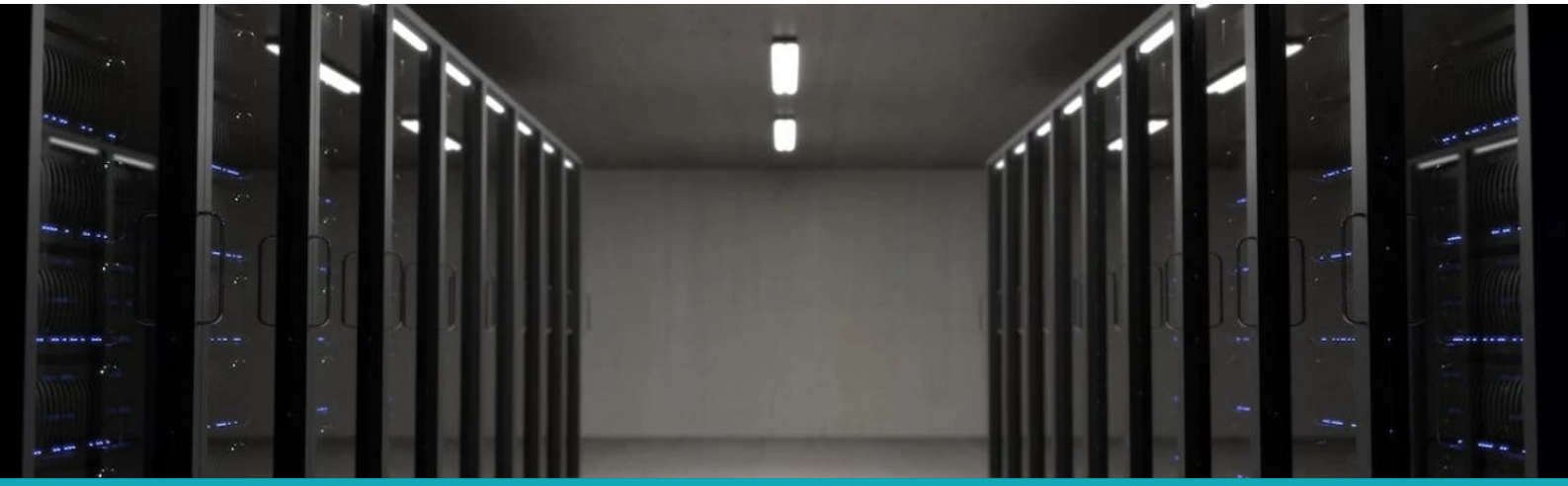
Los datos recopilados de varias fuentes heterogéneas contienen mucho ruido, redundancia y anomalías, lo que aumenta el espacio de almacenamiento al retener datos innecesarios que podrían afectar el flujo de trabajo de administración de datos. Además, los métodos analíticos requieren un cierto nivel de calidad de los datos. Para ello, el preprocesamiento de datos es un paso crucial para garantizar un análisis de datos eficiente. Esta se ocupa de hacer que los datos brutos adquiridos sean aptos para su uso en la toma de decisiones, así como en el uso específico del dominio.

Esta fase puede incluir:

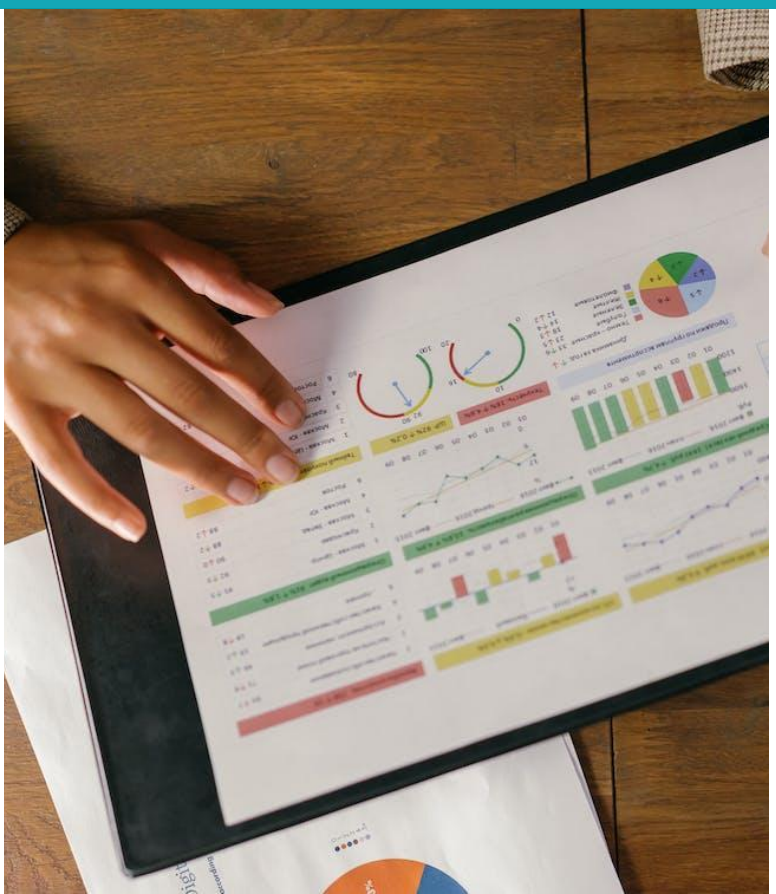
- **Filtrado:** se refiere a la eliminación de datos considerados corruptos de acuerdo con los requisitos de la estrategia de datos de la organización.
- **Extracción:** se refiere a la reelaboración de datos incompatibles, a menudo agrupados o comprimidos específicamente. Esta subfase permite transformar datos dispares en formatos compatibles.
- **Transformación:** se refiere a la modificación, adaptación y empaquetado de datos en formas apropiadas y la estandarización de escalado de atributos para mejorar los procesos de análisis de datos.
- **Validación:** se refiere al establecimiento de reglas de validación y borrado para gestionar las estructuras sintácticas y semánticas de los datos y eliminar datos no válidos y desconocidos.
- **Limpieza:** se refiere a identificar y procesar datos incompletos, inexactos e irrazonables para eliminarlos o completarlos.

Almacenamiento de datos

Se refiere a hacer persistente una gran cantidad de datos recopilados y preprocesados. Las estrategias de los sistemas de almacenamiento tienen un impacto significativo en la escalabilidad y el rendimiento en términos de acceso y exposición de datos.



En la etapa, los sistemas tradicionales como las bases de datos relacionales no se adaptan a estas necesidades y se requieren nuevos sistemas de almacenamiento, ya que se requiere escalabilidad, por lo que se usan sistemas que almacenan los datos de manera distribuida, como es el caso de Hadoop y las bases de datos NoSQL para el almacenamiento de datos modelados.



Otra de las características que debe tener el almacenamiento de datos actual está relacionada con el flujo de los datos ya citados: datos continuos (tiempo real) que se consideran síncronos y con optimización a baja latencia, y asíncrono (los datos se capturan, registran y analizan por lotes).

Análisis de datos

El análisis de datos almacenados utiliza modelos, algoritmos y herramientas adecuadas para proporcionar visibilidad sobre los datos, para que puedan ser consultados en la capa de visualización o capa de consumo.

Una vez que se tienen almacenados los datos, se han de convertir en conocimiento (valor) mediante el análisis de toda la información almacenada.

Se refiere a la manipulación de datos masivos para identificar patrones, encontrar correlaciones y descubrir nuevos modelos de conocimiento emergentes. Esta fase se basa principalmente en capacidades dedicadas de análisis de *Big Data* categorizadas como descriptivas, diagnósticas, predictivas o prescriptivas.

- El análisis **descriptivo** se refiere a la descripción y síntesis de modelos de conocimiento utilizando métodos estadísticos que describen una situación, como informes estándar, cuadros de mando y análisis detallados.

- El análisis **predictivo** se refiere a las probabilidades de predicción empleadas para definir tendencias futuras. Utiliza modelos de aprendizaje supervisados, no supervisados y semisupervisados para proporcionar modelos analíticos predictivos.

- El análisis **prescriptivo** se aplica para predecir eventos futuros e impulsar decisiones proactivas fuera de los límites de la interacción humana.

- **Visualización:** los resultados del análisis de datos es la etapa de consumo de datos que debe permitir su visualización para una correcta toma de decisiones. Esta capa de *Big Data* muestra el producto del almacenamiento y procesamiento de la información, cuyo resultado es la producción de conocimiento. En la actualidad, existe un gran número de herramientas de visualización de datos que proporcionan una gran eficiencia a las compañías. Las herramientas de visualización permiten a los usuarios finales hacer búsquedas y acceder a la información rápidamente, en algunos casos en tiempo real, de modo que los usuarios puedan tener el control de la información en el mismo momento en que se produce.

- **Exposición de datos:** se refiere a la puesta a disposición de los datos para el consumo. Esta exposición consiste en configurar muchas API (interfaces de programación de aplicaciones), respetar las políticas de seguridad y confidencialidad, y permitir el acceso a los datos en diferentes estados: analizados, preprocesados, transformados o incluso tan crudos como se recopilan. La exposición de datos generalmente sirve para muchas aplicaciones internas, como CRM (Customer Relationship Management), para promover productos específicos, pero también podría extenderse para servir a los socios.

A diferencia de como se venían almacenando los datos en las bases de datos tradicionales o en los sistemas de Almacenes de datos, los datos que se almacenan en un lago de datos deben almacenar con ciertas **características particulares**, entre estas tenemos:

- **Datos crudos:** cuando se diseña un sistema *Big Data*, se desea poder responder tantas preguntas como sea posible. Llamaremos coloquialmente a esta propiedad datos crudos. De ser posible, se desea almacenar el detalle más pequeño que se pueda tener a mano.

Por ejemplo, el comercio del mercado de valores es una fuente de información, con millones de acciones y miles de millones de dólares cambiando de manos a diario. Con tantas operaciones, los precios de las acciones se registran históricamente diariamente como precio de apertura, precio alto, precio bajo y precio de cierre. Pero esos *bits* de datos a menudo **no proporcionan una visión general** y pueden sesgar su percepción de lo que sucedió. Por ejemplo, en la figura. Registra los datos de precios de las acciones de Google, Apple y Amazon en un día en que Google anunció nuevos productos dirigidos a sus competidores.

Company	Symbol	Previous	Open	High	Low	Close	Net
Google	GOOG	564.68	567.70	573.99	566.02	569.30	**+4.62**
Apple	AAPL	572.02	575.00	576.74	571.92	574.50	**+2.48**
Amazon	AMZN	225.61	225.01	227.50	223.30	225.62	**+0.01**

Financial reporting promotes daily net change in closing prices. What conclusions would you draw about the impact of Google's announcements?

Fuente: (Marz & Warren, 2015)

Estos datos sugieren que Amazon podría no haberse visto afectado por el anuncio de Google, ya que el precio de sus acciones se movió solo ligeramente. También sugiere que el anuncio no tuvo ningún efecto en Apple o un efecto positivo.

Pero si tiene acceso a los datos almacenados en una granularidad más precisa, puede obtener una imagen más clara de los eventos de ese día e investigar más a fondo las posibles relaciones de causa y efecto.

La siguiente Figura muestra los **cambios relativos minuto a minuto** en los precios de las acciones de las tres compañías, lo que sugiere que tanto Amazon como Apple se vieron realmente afectados por el anuncio, Amazon más que Apple.

Fuente: Acciones de Amazon, Apple y Google. Extraída de Marz y Warren (2015)

También se debe tener en cuenta que los datos adicionales pueden sugerir nuevas ideas que puede no haberse considerado al examinar el resumen original del precio de las acciones. Por ejemplo, los datos más granulares nos hacen preguntarnos si Amazon se vio más afectado porque los nuevos productos de Google compiten con Amazon en los mercados de tabletas y computación en la nube.

El almacenamiento de datos en bruto es muy valioso porque rara vez se conoce de antemano todas las preguntas que se desea responder. Al mantener los datos más crudos posibles, maximiza su capacidad de obtener nuevos conocimientos, mientras que resumir, sobrescribir o eliminar información limita lo que los datos pueden decirnos.

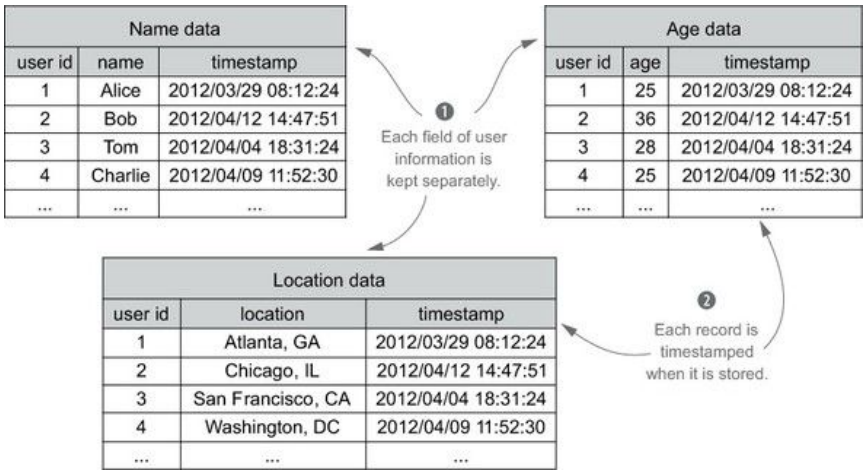
Aunque esto implica más almacenamiento, las tecnologías *Big Data* están diseñadas para administrar *petabytes* y *exabytes* de datos. Específicamente, administran el almacenamiento de sus datos de manera distribuida y escalable, al tiempo que admiten la capacidad de consultar directamente los datos.

- La segunda característica de un lago de datos es que los **datos son inmutables:** pueden parecer un concepto extraño si trabaja con las bases de datos relacionales. Después de todo, en el mundo de las bases de datos relacionales, y también en la mayoría de las otras bases de datos, la actualización es una de las operaciones fundamentales. Pero para la inmutabilidad no se actualiza o elimina datos, solo se agrega una nueva versión del dato. En *big data* se obtienen dos ventajas.

- **Tolerancia a fallas humanas:** las personas cometen errores, se debe limitar el impacto de tales errores y tener mecanismos para recuperarse de ellos. Con un modelo de datos mutables, un error puede causar la pérdida de datos, porque los valores se borran en la base de datos. Con un modelo de datos inmutable, no se pierden datos. Si se escriben datos incorrectos, aún existen versiones anteriores de los datos que son correctas. La reparación del sistema de datos es solo una cuestión de eliminar las versiones de datos defectuosas y volver a calcular las vistas creadas a partir del conjunto de datos maestros.

- **Simplicidad:** los modelos de datos mutables implican que los datos deben indicarse de alguna manera para poder recuperar y actualizar objetos de datos específicos. En contraste, con un modelo de datos inmutable solo necesita la capacidad de agregar nuevas versiones de datos al conjunto de datos maestro. Esto no requiere un índice y es tan simple como usar archivos planos.

Esta característica es posible debido a que en un esquema inmutable cada dato se almacena por separado como se observa en la siguiente figura que muestra el nombre, la edad y la dirección de los usuarios.



Fuente: Simplicidad de datos. Extraída de Marz y Warren (2015)

Si el usuario 3, llamado Tom, se muda a Los Ángeles, solo se agregará la nueva versión a la columna correspondiente y se mantendrán las versiones anteriores.

La ubicación actual de Tom implica una consulta simple sobre los datos: observar las ubicaciones y elegir la que tenga **la marca de tiempo más reciente**. Al mantener cada campo en una tabla separada, solo se registra la información que cambió. Esto requiere menos espacio para el almacenamiento y garantiza que cada registro sea información nueva y no se transfiera simplemente del último registro.

Location data		
user id	location	timestamp
1	Atlanta, GA	2012/03/29 08:12:24
2	Chicago, IL	2012/04/12 14:47:51
3	San Francisco, CA	2012/04/04 18:31:24
4	Washington, DC	2012/04/09 11:52:30
3	**Los Angeles, CA**	**2012/06/17 20:09:48**
...	...	...

1. The initial information provided by Tom (user id 3), timestamped when he first joined FaceSpace.
2. When Tom later moves to a new location, you add an additional record timestamped by when you received the new data.

Fuente: Ejemplo de datos. Extraída de Marz y Warren (2015)

En este caso si este cambio de dirección es errado solo se requiere borrar la última versión.

Location data		
user id	location	timestamp
1	Atlanta, GA	2012/03/29 08:12:24
2	Chicago, IL	2012/04/12 14:47:51
3	San Francisco, CA	2012/04/04 18:31:24
4	Washington, DC	2012/04/09 11:52:30
~~3~~	~~Los Angeles, CA~~	~~2012/06/17 20:09:48~~
...	...	...

Human faults can easily be corrected by simply deleting erroneous facts. The record is automatically reset by using earlier timestamps.

Fuente: Ejemplo de datos. Extraída de Marz y Warren (2015)

Datos eternamente verdaderos:

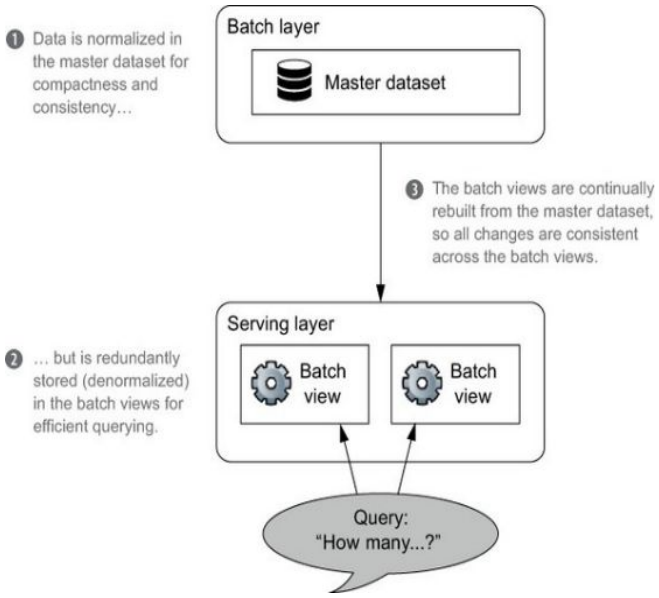
- La consecuencia clave de la inmutabilidad es que cada dato es verdadero a perpetuidad. Es decir, un dato, una vez verdadero, siempre debe ser verdadero. La inmutabilidad no tendría sentido sin esta propiedad. Etiquetar cada unidad de datos con una marca de tiempo es una forma práctica de hacer que los datos sean eternamente verdaderos.

Esta mentalidad es la misma que cuando aprendemos historia en la escuela. Por ejemplo Venezuela tenía 4 provincias en el año 1800, en ese momento ese dato era verdadero aunque ahora sea diferente la división territorial de ese país.

El modo de flujo de datos identificado en la etapa de adquisición determina las técnicas de procesamiento de datos lo que a su vez impacta en la arquitectura de las soluciones de *big data.*

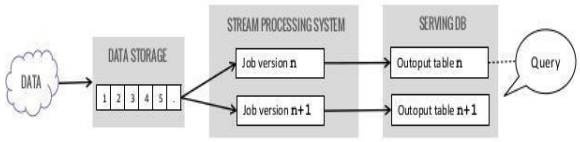
Entre las técnicas de procesamiento de datos se tienen:

- **El paradigma Batch** (flujo de datos por lotes) consiste en procesar volúmenes de datos en tiempos espaciados, es decir, cada cierto tiempo. Esto es posible ya que el sistema almacena en lotes toda la información que va obteniendo por cada periodo.
  Esta técnica requiere de una capa de lote donde se almacena toda la data y una capa de servicio donde se definen las vistas que permiten hacer las consultas, este tiene como características que es escalable (Volumen de datos), procesa grandes cantidades de información estática con una alta latencia.



Fuente: CITATION Nat15 \| 3082 (Marz & Warren, 2015)

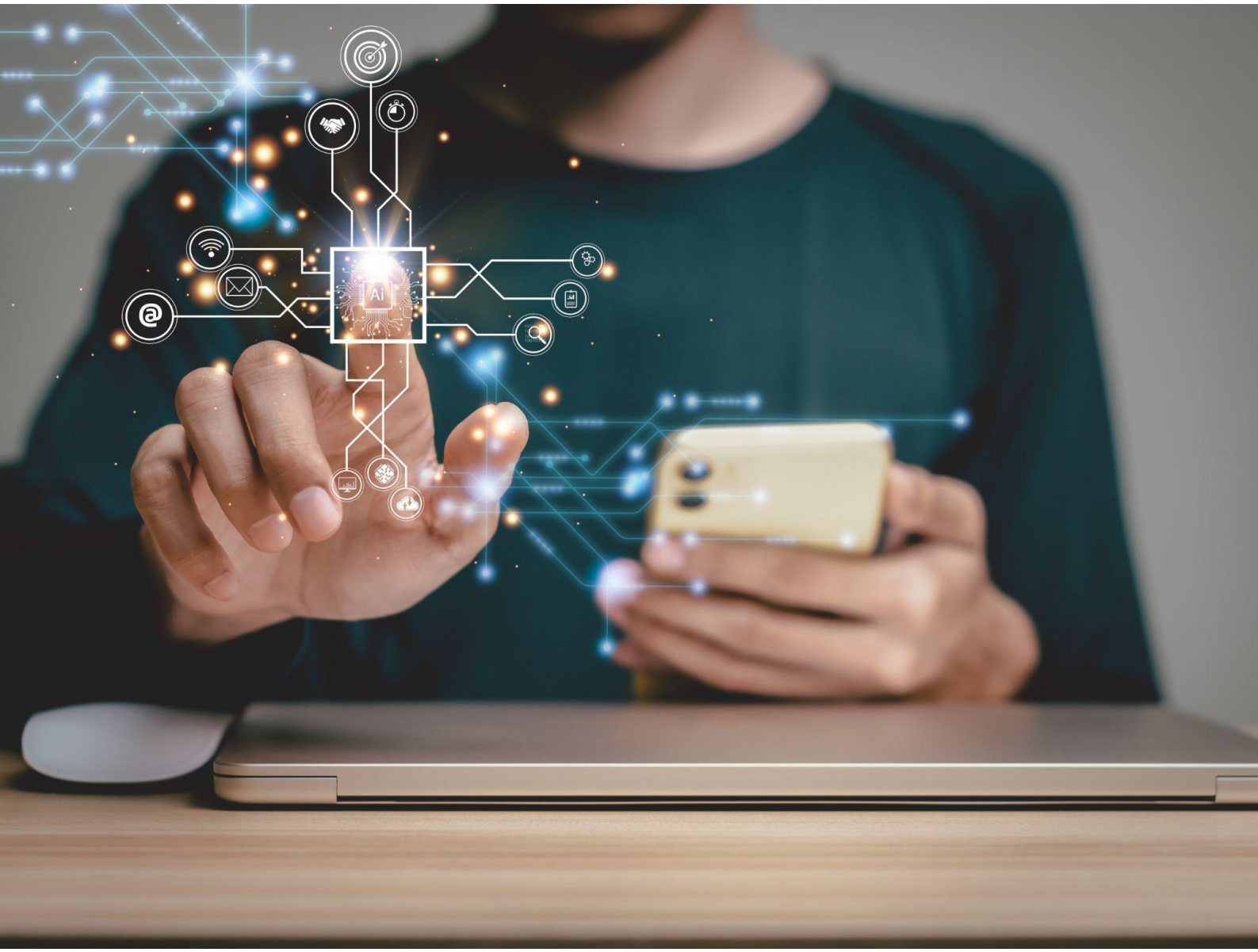
- **El paradigma straming** (flujo de datos continuos) que consiste en procesar volúmenes de datos en tiempos lo más parecido al tiempo real, nos referimos a órdenes de 100 milisegundos a segundos. Para ello los datos se procesan a medida que se reciben, lo cual garantiza una baja latencia en el procesamiento de los datos pero los resultados de las consultas son menos precisos que en el caso del paradigma batch ya que se cuenta con datos parciales, pero vigentes, por lo que cumple con la propiedad de Velocidad de *Big Data.*



De estos dos paradigmas surgen técnicas de procesamiento híbridas como Lambda y Kappa las cuales serán asignadas a una actividad de evaluación.

En este contenido pudimos comparar ***Big data*** con áreas de conocimiento relacionadas como son Inteligencia de Negocio o business inteligence y la Análitica de negocio o business analytic, para luego observar la evolución de la arquitectura de Inteligencia de Negocios a *Big Data.*

En segundo lugar, conocimos las etapas de la cadena de valor de *big data*, las características de un lago de datos y los principales paradigmas de procesamiento de datos como los son el de lotes y el de flujo continuo de datos.



Aguilar, J. (2019). *Inteligencia de negocios y analítica de datos.* Alfaomega Grupo Editor, S.A. de C.V.

Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*, pp. 3-9.

Díaz, J. (2011). *Introducción al Business Intelligence.* UOC.

Faroukhi, A., Alaoui, I., Youssef, G. y Amine, A. (2020). A Multi-Layer Big Data Value Chain Approach for Security Issues. *Procedia Computer Science, 175*, pp. 737-744. https://www.researchgate.net/publication/343498016_A_Multi-Layer_Big_Data_Value_Chain_Approach_for_Security_Issues

Luhn, H. (1958). A Business Intelligence System. *IMB Journal, 2*(4), pp. 314-319. https://doi.org/10.1147/rd.24.0314

Porter, M. (1985). *Competitive Advantage. Creating and Sustaining Superior Performance.* Free Press.

Referencias de las imágenes

Big Data International Campus (2018). Diferencias entre *big data, business intelligence* y *business analytics* [Imagen]. Disponible en: https://www.campusbigdata.com/big-data-blog/item/148-diferencias-entre-big-data-business-analytics-y-business-intelligence

InnoWiki (2014). Business Intelligence [Imagen]. Disponible en: http://185.5.126.23/innowiki/index.php/Business_Intelligence

Jony, R., Rony, R. y Rahman, M. (2016). The Data-Information-Knowledge-Wisdom hierarchy pyramid [Imagen]. Disponible en: *Big Data Characteristics, Value Chain and Challenges*. https://www.researchgate.net/profile/Musfiqur-Rahman/publication/311323742_Big_Data_Characteristics_Value_Chain_and_Challenges/links/5841ae8308ae8e63e6219215/Big-Data-Characteristics-Value-Chain-and-Challenges.pdf

Marz, N. y Warren, J. (2015). Acciones de Amazon, Apple y Google [Imagen]. Disponible en: *Big Data Principles and best practices of scalable real-time data systems*. Manning Publications Co.

Marz, N. y Warren, J. (2015). Citation [Imagen]. Disponible en: *Big Data Principles and best practices of scalable real-time data systems.* Manning Publications Co.

Marz, N. y Warren, J. (2015). Ejemplo de datos [Imagen]. Disponible en: *Big Data Principles and best practices of scalable real-time data systems.* Manning Publications Co.

Marz, N. y Warren, J. (2015). Ejemplo de datos crudos [Imagen]. Disponible en: *Big Data Principles and best practices of scalable real-time data systems.* Manning Publications Co.

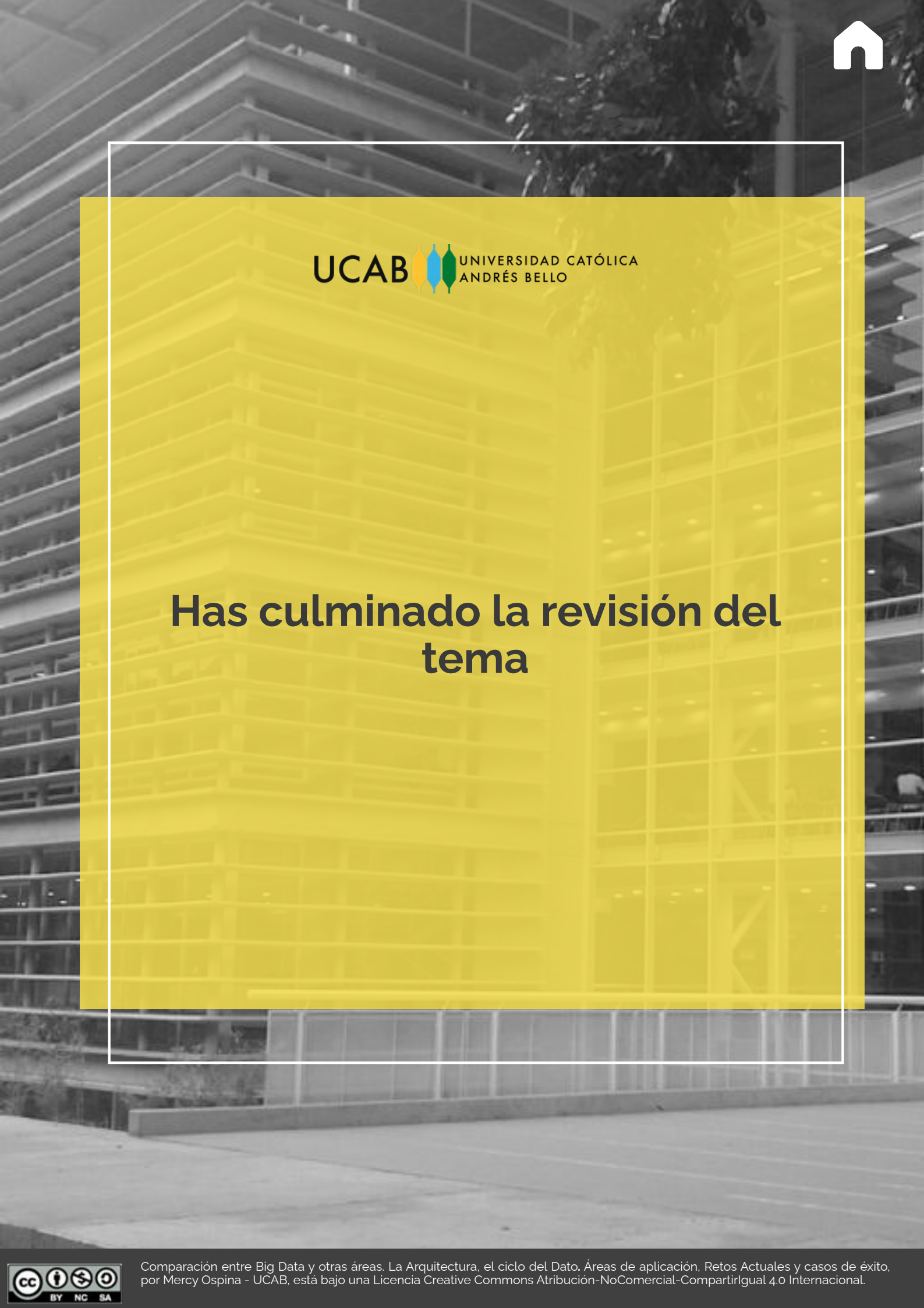
Marz, N. y Warren, J. (2015). Paradigma straming [Imagen]. Disponible en: *Big Data Principles and best practices of scalable real-time data systems.* Manning Publications Co.

Marz, N. y Warren, J. (2015). Simplicidad de datos [Imagen]. Disponible en: *Big Data Principles and best practices of scalable real-time data systems*. Manning Publications Co.

PNGWING (s.f.). Arquitectura de datos [Imagen]. Disponible en: https://www.pngwing.com/es/free-png-hjzyz

Bibliografía sugerida

Pérez, G. (2020). *¿Business Intelligence o Business Analytics?* Not Just BI. https://notjustbi.com/business-intelligence-o-business-analytics/



UCAB UNIVERSIDAD CATÓLICA ANDRÉS BELLO

Has culminado la revisión del tema