Unidad 2: Introducción a Big Data

2.1 Historia de BD
2.2 Definición de Big Data
2.3. Procesamiento de Big Data
2.4. Componentes de una Arquitectura de Big Data.
2.5 Roles
2.6 Uso ético del Big Data, principios de privacidad de GDMA

2.1 Historia de Big Data

Historia del Big Data

La historia del Big Data es muy poco conocida. Como sucede frecuentemente con las tendencias, parece que cuando explotan es algo muy novedoso que acaba de aparecer; pero en muchos casos son la eclosión de algo que ha estado madurando durante largo tiempo.

Como hemos comentado en otras ocasiones, el Big Data es el análisis de un gran volumen de conjuntos de datos. Para recolectar, tratar y analizar esa ingente cantidad de información se necesitan fórmulas de procesamiento potentes y rápidas. Por ello, estas técnicas parecen haber surgido recientemente, gracias a los avances tecnológicos.

Esto, en gran medida es así: el término se comienza a utilizar de forma generalizada a finales de los años 90 y el boom llega con los avances experimentados en campos como internet, dispositivos móviles y conexión. Sin embargo, la historia del Big Data se remonta a mucho antes.

Hay incluso quienes lo sitúan en el paleolítico, con una lógica que relaciona el término con el primitivo interés de los seres humanos por lograr y procesar la información. A continuación, ahondamos un poco más en esta cuestión compartiendo la historia del Big Data:

- En el Paleolítico Superior se empleaban rudimentarios métodos de almacenamiento de datos con el empleo de palos o muescas en huesos. Con este sistema, se podía llevar cuenta de provisiones, realizar cálculos básicos e incluso predecir necesidades de comida para el grupo. Quizá sea demasiado incluirlo en la historia del big data, pero es el primer momento documentado en el que la humanidad se interesa por los datos: el germen de todo lo que viene después. Si bien las cantidades no podían ser muy grandes, es la primera evidencia del interés por recopilar, contar y guardar datos.
- 2400 AC. En Babilonia se extiende el uso del ábaco, un sistema para realizar cálculos. En esta época surgen también las primeras bibliotecas como lugares para almacenar y consultar conocimiento.
- 48 AC. Los Romanos invaden Alejandría y accidentalmente destruyen su famosa biblioteca. Parte de los fondos se trasladaron a otros lugares, pero la mayoría de la colección fue quemada, perdida o robada. Hasta el momento, había logrado reunir medio millón de documentos con la intención de almacenar todo el conocimiento de la Humanidad.
- Siglo II AC. Se desarrolla la primera computadora mecánica conocida en Grecia. El mecanismo de Anticitera era un aparato analógico de bronce diseñado para predecir posiciones astronómicas, probablemente una

evolución de otros sistemas que no se han recuperado a día de hoy. Se empleó para el estudio astrológico y para marcar el calendario, fundamentalmente las fechas exactas de los antiguos Juegos griegos.

- 1663. John Graunt realiza el primer experimento de análisis de datos estadísticos conocido. Con los datos de funciones, teoriza un sistema de alerta para la peste bubónica en toda Europa.
- 1792. Aunque hay constancia de análisis estadísticos desde las Guerras del Peloponeso y la palabra estadística se acuña en Alemania unos años antes; en 1792 se asocia el término a la “colección y clasificación de datos“.
- 1865. Aparece por primera vez el término business intelligence, en la enciclopedia comercial de Richard Millar Devens. En ella describe cómo el banquero Henry Furnese logró una importante ventaja competitiva recogiendo, estructurando y analizando datos clave de su actividad. La inteligencia de negocio es sin duda uno de los grandes motores de la analítica dentro de la historia del big data.
- 1880. Herman Hollerith, empleado del censo estadounidense, desarrolla su máquina tabuladora. Con ella consigue reducir un trabajo de 10 años a 3 meses. Este ingeniero funda una compañía que posteriormente se conocería como IBM.
- 1926. Nikola Tesla predice la tecnología inalámbrica. Según su visión, el planeta en un gran cerebro en el que todo está conectado, por lo que deberíamos ser capaces simplificar el uso del teléfono. Predice que cada hombre llevará uno en su propio bolsillo.
- 1928. El ingeniero alemán Fritz Pfleumer patenta el primer sistema magnético para almacenar datos. Sus principios de funcionamiento se utilizan hoy en día.
- 1944. Primer intento de conocer la cantidad información que se crea. Se trata de un estudio académico de Fremont Rider, que pronostica 200 millones de libros en la Universidad de Yale en 2040, almacenados 6.000 millas de estanterías.
- 1958. El informático alemán Hans Peter Luhn, define la inteligencia de negocio: la habilidad de percibir las interrelaciones de los hechos presentados para guiar acciones hacia un objetivo deseado. En 1941 pasó a ser Gerente de Recuperación de Información en IBM.
- 1962. Se presenta IBM Shoebox en la Expo de 1962. Creada por William C. Dersch supone el primer paso en el reconocimiento de voz, capaz de registrar palabras en inglés en formato digital.
- 1965. Se proyecta el primer data center en Estados Unidos, para guardar documentación de impuestos y huellas dactilares en cintas magnéticas. Un año antes comienzan a surgir voces que alertan del problema de guardar la ingente cantidad de datos generada.
- 1970. IBM desarrolla el modelo relacional de base de datos, gracias al matemático Edgar F. Codd. Este científico inglés es también responsable de las doce leyes del procesamiento analítico informático y acuño el término OLAP.
- 1976. Se populariza el uso de MRP (software de gestión de materiales), antecedentes de los ERP actuales, que mejoran la eficiencia de las operaciones en la empresa; además de generar, almacenar y distribuir datos en toda la organización.

- 1989. Erik Larson habla por primera vez de Big Data en el sentido que conocemos la expresión hoy en día. La revista Harpers Magazine recoge su artículo, en el que especula sobre el origen del correo basura que recibe. En torno a este año se empiezan a popularizar las herramientas de business intelligence para analizar la actividad comercial y el rendimiento de las operaciones.
- 1991. Nace internet, a la postre, la gran revolución de la recolección, almacenamiento y análisis de datos. Tim Berners-Lee establece las especificaciones de un sistema de red con interconexiones a nivel mundial accesible para todos en cualquier lugar.
- 1993. Se funda QlikTech, germen de la actual Qlik, que crea un sistema revolucionario de business intelligence (en 2012, Gartner comienza a hablar de business discovery para definir ese tipo de análisis).
- 1996. Los precios del almacenamiento de datos empiezan a ser accesibles con un coste eficiente en lo que es una de las grandes revoluciones en la historia del big data. El libro **La evolución de los sistemas de almacenamiento**, de 2003, establece esta fecha como el primer año en el que el almacenamiento digital es más barato que el papel.
- 1997. Google lanza su sistema de búsqueda en internet y en los siguientes años será de largo el primer lugar al que acudir en busca de datos en internet. Este mismo año, se publica el estudio ¿Cuánta información hay en el mundo?, de Michael Lesk. La conclusión es que hay tanta y crece a tal velocidad, que gran parte de ella no será vista por nadie jamás.
- 1999. El término Big Data es analizado por primera vez en un estudio académico. La Asociación de Sistemas Informáticos recomienda centrarse en el análisis de información ya que existe gran cantidad de datos y no todos son útiles. Recuerdan el propósito de la computación, que es el entendimiento, no los números.
- 2001. Doug Laney, de Gartner, define las 3 V's del Big Data. Este es un hito clave en la historia del big data. Se trata de tres conceptos que definen el término: volumen, velocidad y variedad. Al mismo tiempo, se populariza el concepto SaaS (software as a service).
- 2005. Nace la Web 2.0, una web donde predomina el contenido creado por los usuarios. Este mismo año se crea Hadoop, un entorno de trabajo Big Data de software libre.
- 2007. La revista Wired publica un artículo que lleva el concepto de Big Data a las masas.
- 2010. Los datos que se generan en dos días equivalen a la cantidad de datos generados desde el inicio de la civilización hasta 2003, según Eric Schmidt (Google).
- 2013. El archivo de mensajes públicos de Twitter en la Biblioteca del Congreso de Estados Unidos llega a los 170 billones de mensajes, creciendo a ritmo de 500 millones al día. Según la institución que alberga algunos de los documentos históricos más importantes del mundo, dicho archivo ofrece una imagen más amplia de las normas culturales, diálogos, tendencias y eventos de hoy en día. De este modo, contribuye a una mejora de la información en procesos legislativos, educación, definición de autoría de nuevos trabajos y otras cuestiones.

- 2014. Los móviles superan a los ordenadores en accesos a internet. La conexión casi continua contribuye a generar muchos más datos y mejora la conectividad con otros dispositivos.
- 2016. El Big Data se convierte en la palabra de moda. Se generaliza la contratación de expertos en Big Data, el Machine Learning llega a las fábricas y el Internet de las Cosas empieza a impregnarlo todo.
- 2017. Los datos llegan a las masas. La gente controla sus patrones de descanso con pulseras, sabe en qué se gasta el dinero con aplicaciones móviles y se informa sobre la posesión de balón de su equipo de fútbol. Los datos están en todas partes y la población está ya predispuesta a usarlos.

Futuro. ¿Qué nos deparará el futuro? Muy difícil de pronosticar, pero seguramente un aumento de datos y la consiguiente necesidad de tecnología para recogerlos, adaptarlos, almacenarlos y analizarlos. La computación cuántica está a la vuelta de la esquina y la historia del big data sigue avanzando.

2.2 Definición de BIG DATA

Big Data es un término que describe un gran volumen de datos, tanto estructurados como no estructurados.

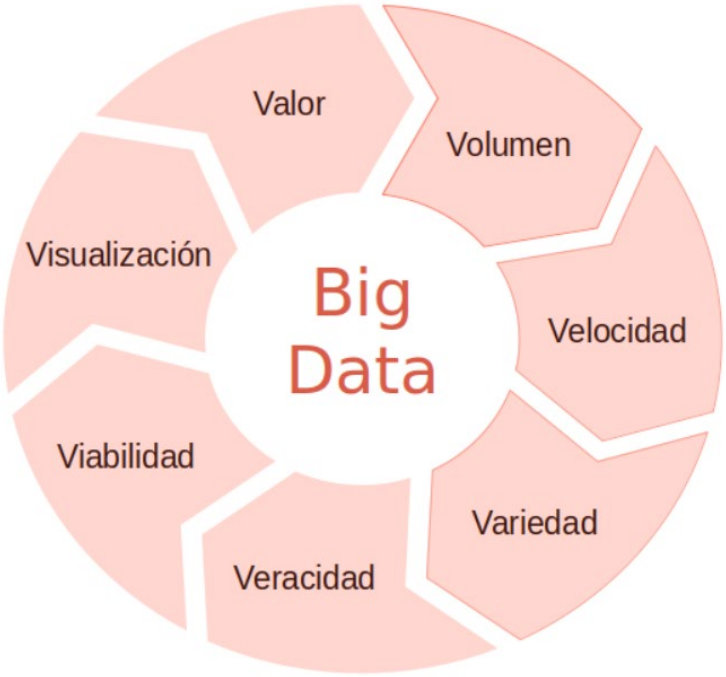
La cantidad de datos es tan grande, que las aplicaciones de software de procesamiento de datos que tradicionalmente se venían usando no son capaces de capturar, tratar y ponerlos en valor en un tiempo razonable.

No toda la información es Big Data. Para ser considerada como tal, debe cumplir con las llamadas las 7 Vs.

Las 7 Vs

Dijimos que Big Data es un concepto que se refiere a grandes volúmenes de datos que son muy variados y veloces, al punto de que resulta muy complicado capturarlos y procesarlos con métodos tradicionales.

Para hablar de Big Data, necesitamos tener en cuenta las características que deben cumplir los datos para ser considerados Big Data



Volumen: Es la cantidad de datos que son generados y se almacenan con la finalidad de procesarlos para transformar los datos en acciones. Es la característica más asociada al Big Data.

Velocidad: Es la rapidez con la que los datos son creados, almacenados y procesados en tiempo real. Considera la frecuencia con la que se generan los datos, el tiempo de análisis y el de espera para que la información se encuentre disponible.

Variedad: Es la forma, el tipo y la fuente de los datos. Pueden ser:

- Estructurados y no estructurados.
- Públicos, privados, comunitarios, etc.
- Documentos, correos, multimedia, redes sociales, etc.

Veracidad: Es el alto grado de fiabilidad, integridad y autenticidad de los datos (incertidumbre de los datos).

Viabilidad: Capacidad de generar un uso eficaz del gran volumen de datos que se manejan. Implica:

- Respeto a la privacidad y la confidencialidad.
- Personalización de servicios para su uso eficaz.

Visualización: Es el modo en que los datos son presentados para encontrar patrones y claves ocultas en el tema a investigar. Pretende una representación de resultados complejos en un formato sencillo e interactivo con los usuarios.

Valor: Los datos que se transforman en información, y luego en conocimiento para poder tomar una acción, decisión o diseñar una estrategia. Supone la combinación de información, contexto y sentido.

Poder prever el futuro o plantear los posibles escenarios, permite que estemos preparados para entender lo que puede pasar y tomar decisiones por adelantado. Una de las formas que tiene el análisis de datos para hacer previsiones sobre el futuro, es el estudio de lo que ha pasado y los datos que ya tenemos registrados.

No es la cantidad de datos lo que es importante. Lo que importa con el Big Data es lo que las organizaciones hacen con los datos.

2.3 Procesamiento de Big Data

La Big Data se procesa dividiéndola en partes más pequeñas, aplicando tecnologías como Spark, Hadoop y servicios de cómputo en la nube (Amazon Web Services, Google Cloud Platform, Microsoft Azure)

Los datos son vitales para las empresas de hoy. Con ellos se toman decisiones y se crean mejores productos.

Los objetivos de Data Science son:

- ✓ Tomar decisiones y crear estrategias de negocio para sacar el máximo potencial de una empresa.
- ✓ Crear productos de software más inteligentes y funcionales.

¿De qué trata este proceso?

Los pasos a seguir son los siguientes:

1. Obtener los datos: (mediciones directas, encuestas, internet)
2. Transformar y limpiar los datos (incompletos o formato incorrecto)
3. Explorar, analizar y visualizar los datos (patrones, tendencias, insights para presentar en visualizaciones o reportes amigables)
4. Usar modelos de Machine Learning (IA): predecir información
5. Integrar datos e IA a productos de software: (escalar estos modelos para ponerlos a disposición del usuario final.)

Los tres primeros pasos dependen de información histórica, y los dos últimos se basan en predicciones.

2.4 Componentes de una arquitectura de big data

1. Fuentes de Datos

- Datos Estructurados: Bases de datos relacionales (RDBMS), como MySQL,

PostgreSQL.
- Datos Semiestructurados: Archivos XML, JSON, logs de servidor.
- Datos No Estructurados: Documentos, imágenes, videos, archivos de audio, correos electrónicos.
- Datos de Transacciones Datos generados por aplicaciones transaccionales.
- Datos de Streaming: Datos en tiempo real provenientes de sensores, redes sociales, dispositivos IoT.

2. Ingesta de Datos
- Batch Processing: Herramientas como Apache Hadoop para procesamiento por lotes.
- Stream Processing: Herramientas como Apache Kafka, Apache Flink, Apache Storm para procesamiento en tiempo real.

3. Almacenamiento de Datos
- Data Lakes: Utilizando tecnologías como HDFS (Hadoop Distributed File System), Amazon S3, Azure Data Lake.
- Data Warehouses: Utilizando tecnologías como Amazon Redshift, Google BigQuery, Snowflake.
- Bases de Datos NoSQL: Bases de datos como MongoDB, Cassandra, HBase.
- Bases de Datos SQL: Bases de datos relacionales tradicionales.

4. Procesamiento y Análisis de Datos
- MapReduce: Un modelo de programación y su implementación asociada en Apache Hadoop.
- Motor de Procesamiento de Datos: Herramientas como Apache Spark para procesamiento distribuido.
- Machine Learning y Data Mining: Herramientas y bibliotecas como TensorFlow, PyTorch, Scikit-learn, Apache Mahout.

5. Acceso y Consulta de Datos
- Lenguajes de Consulta: SQL, HiveQL para consultas de datos estructurados.
- Herramientas de BI y Visualización: Tableau, Power BI, Looker, QlikView para visualización y reporting.
- APIs y Servicios Web: Para acceder y consultar datos desde aplicaciones externas.

6. Gestión de Datos
- Data Governance: Políticas y procedimientos para asegurar la calidad y seguridad de los datos.
- Metadata Management: Gestión de metadatos para entender la procedencia y contexto de los datos.
- Data Lineage: Seguimiento del flujo de datos a través de su ciclo de vida.

7. Seguridad de Datos
- Encriptación de Datos: En reposo y en tránsito.

- Control de Acceso: Autenticación y autorización de usuarios y aplicaciones.
- Auditoría y Monitoreo: Registro y monitoreo de acceso a los datos.

8. Infraestructura y Plataformas

- On-premise: Infraestructura local en los servidores de la empresa.
- Cloud: Plataformas en la nube como Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure.
- Híbrido: Combinación de infraestructura on-premise y cloud.

Ejemplo de Arquitectura de Big Data

Un ejemplo de cómo estos componentes pueden trabajar juntos:

1. Ingesta de Datos: Datos de sensores IoT y redes sociales se capturan en tiempo real utilizando Apache Kafka.
2. Almacenamiento de Datos: Los datos se almacenan en un Data Lake utilizando Amazon S3.
3. Procesamiento de Datos: Apache Spark se utiliza para procesar los datos y realizar análisis en batch y en tiempo real.
4. Análisis Avanzado: Modelos de machine learning se entrenan utilizando TensorFlow y los resultados se almacenan en un Data Warehouse (Amazon Redshift).
5. Visualización: Power BI se utiliza para crear dashboards interactivos que permiten a los usuarios finales consultar y visualizar los datos analizados.
6. Seguridad y Gobernanza: Los datos se encriptan en tránsito y en reposo, con políticas de control de acceso implementadas para asegurar que solo los usuarios autorizados puedan acceder a los datos.

Esta combinación de componentes permite una gestión y análisis eficiente de grandes volúmenes de datos, proporcionando insights valiosos y soporte para la toma de decisiones basada en datos.

ID Libro	Título	Autor	Editorial	Género	ISBN
001	Justicia Auxiliar	Ann Leckie	Nova	Ciencia Ficción	xxxxxxxxx
002	La ciudad que nos unió	N.K. Jemesin	Nova	Ciencia Ficción	xxxxxxxxx
003	La Historia Interminable	Michael Ende	Santillana	Fantasía	xxxxxxxxx
004	Sakura	Matilde Asensi	Esfera de los libros	Ficción/Suspense	xxxxxxxxx
005	Neverwhere	Neil Gaiman	Roca Libros	Fantasía Urbana	xxxxxxxxx

Estructurados

Los datos estructurados suelen llamar también datos internos cuantitativos, y es el tipo de datos con el que la mayoría de nosotros estamos acostumbrados a trabajar. Son datos que encajan perfectamente en campos y columnas fijos en bases de datos y hojas de cálculo.

Los ejemplos de datos estructurados incluyen nombres, fechas, direcciones, números de tarjetas de crédito, información bursátil, etc.

Los datos estructurados están muy organizados y la computadora es capaz de comprenderla fácilmente. Quienes trabajan con bases de datos relacionales pueden ingresar, buscar y manipular datos estructurados con relativa rapidez. Esta es la característica más atractiva de los datos estructurados.

Semi-estructurados

Los datos semiestructurados son los datos que no se ajustan a un modelo de datos pero tienen alguna estructura. Carecen de un esquema fijo o rígido. Son los datos que no están en una base de datos relacional pero que tienen algunas propiedades organizativas que facilitan su análisis. Con algún proceso, podemos almacenarlos en la base de datos relacional.

No estructurados

Los datos no estructurados se los conoce también como datos internos cualitativos y no pueden procesarse y analizarse utilizando herramientas y métodos convencionales.

Los ejemplos de datos no estructurados incluyen texto, video, audio, actividad móvil, actividad en redes sociales, imágenes satelitales, imágenes de vigilancia, etc.

Los datos no estructurados son difíciles de deconstruir porque no tienen un modelo predefinido, lo que significa que no se pueden organizar en bases de datos relacionales.

Fuentes de Datos

Los datos internos

Los datos internos son información, estadísticas y tendencias que las organizaciones descubren a través de sus operaciones.
Incluye hechos y cifras que las empresas obtienen de bases de datos internas, software, clientes e informes.
También podemos definir datos internos como información creada por la operación de una organización que incluye ventas , órdenes de compra y transacciones en el inventario Este concepto se contrapone al de los datos creados por un estudio o base de datos independiente.
Los datos internos son datos recuperados desde dentro de la empresa para tomar decisiones para operaciones exitosas. Esta información es importante para determinar si las estrategias que la empresa está utilizando actualmente son acertadas o si se deben hacer cambios.
Hay cuatro áreas diferentes de las que una empresa puede recopilar datos internos: ventas, finanzas, marketing y recursos humanos. Cada área proporciona una perspectiva única, pero los datos conectan los departamentos.

Datos Departamento de Ventas

El departamento de ventas es esencial para la rentabilidad de una empresa.
Los datos de ventas pueden incluir ingresos, rentabilidad, canales de distribución, puntos de precio, perfiles de cliente y las brechas entre lo que se produce y lo que compran los clientes. Los datos de ventas pueden ayudar a los propietarios de negocios a comprender las áreas de fortaleza y las áreas de debilidad, que pueden impulsar un cambio en el marketing o el enfoque.

Datos Departamento de finanzas

El departamento de finanzas de una empresa puede generar datos valiosos como informes de producción, informes de flujo de caja y presupuestos.
Los informes de producción detallan las cantidades exactas gastadas para fabricar productos y servicios. Los informes de flujo de caja detallan cuánto dinero se utilizó dentro de la empresa durante un período de tiempo específico. Los presupuestos proporcionan información sobre cómo se gastó el dinero en relación con lo que se asignó.
A diferencia de las ventas, que brindan información sobre la cantidad de productos o servicios vendidos, los datos financieros revelan lo que gasta una empresa para fabricar estos productos y servicios, y la variación en estos costos. Por ejemplo, un informe financiero puede mostrar que pedir ciertos suministros es más barato en verano que en invierno.

Datos Departamento de marketing

El departamento de marketing de una organización se centra en promover productos y servicios, crear conciencia de marca y dirigirse adecuadamente a clientes y posibles clientes.

Los departamentos de marketing pueden generar informes sobre el comportamiento del cliente, los perfiles de los clientes, la cantidad de campañas en las redes sociales, el nivel de conocimiento de la marca, el nivel de participación de mercado en relación con la competencia y el nivel de participación a través del sitio web y el contenido.
El análisis del sistema de datos internos de marketing interno puede ayudar a los propietarios de empresas a decidir qué campañas de marketing están funcionando, cuáles necesitan mejoras y qué tipo de nuevas campañas serían eficaces en función de las necesidades de los consumidores que se desean obtener.

Datos Departamento de RRHH

Los departamentos de recursos humanos pueden proporcionar información sobre lo que cuesta contratar y capacitar a un empleado, la productividad de un empleado individual, cómo el ausentismo está afectando la cultura laboral y el nivel de satisfacción o insatisfacción de los empleados con respecto a la empresa.
Una empresa no puede prosperar si los empleados no están contentos, son improductivos y están desmotivados. Los datos de recursos humanos pueden revelar las áreas en las que una empresa necesita mejorar sus procesos para que los trabajadores se sientan empoderados y valorados y, por lo tanto, es más probable que se comprometan a colaborar con sus habilidades, talento y esfuerzo.

Protección de los datos internos de la empresa

Una de las mayores amenazas para los datos internos de la empresa puede ser sus propios empleados. Ya sea accidentalmente o debido a intenciones maliciosas, el resultado final es igualmente peligroso. Por eso es crucial que se implementen políticas de seguridad de datos internos.
Estas son las 4 prácticas recomendadas para proteger la información de la empresa.

- **Evaluar la situación**
  Se aconseja a las organizaciones que realicen evaluaciones de riesgos en forma regular. Si los miembros del personal no tienen las habilidades adecuadas, debe considerarse el contratar una empresa externa. Esto aliviaría al personal técnico interno y revelaría posibles agujeros de seguridad, como la capacidad de acceder a la red interna a través de dispositivos de terceros o aplicaciones basadas en la nube, que los procesos internos no hayan descubierto.

- **Autenticación en dos factores**
  En este tipo de proceso, se solicita a los usuarios que proporcionen las credenciales habituales (como su identificación de empleado y contraseña), así como un código de un solo uso, que normalmente se envía a sus dispositivos móviles.

  Este tipo de modelo de autenticación proporciona a las Pymes diferentes beneficios. En primer lugar, este proceso dificulta que los empleados compartan las credenciales de inicio de sesión, lo que a su vez reduce las

posibilidades de un compromiso accidental si los empleados olvidan cerrar la sesión o apagar sus computadoras.
La autenticación de dos factores también facilita el seguimiento de los intentos de inicio de sesión específicos, lo que permite que el proceso de investigación y reparación de posibles fraudes de datos consuma mucho menos tiempo.

- **Concienciación de los empleados**
  Todo el equipo, desde la administración hasta la equipo técnico y los empleados de primera línea, debe asumir la responsabilidad de evitar correos electrónicos no deseados, cambiar regularmente sus contraseñas y descargar solo aplicaciones de terceros aprobadas en dispositivos móviles y estaciones de trabajo.
  La mejor opción es restringir el acceso a la información a solamente una persona y sólo a la que necesita ver para poder trabajar en una tarea específica. Cuando los empleados completan un proyecto o ya no están asignados al equipo asociado, se debería cambiar su perfil de acceso de inmediato.
  Lo mismo ocurre con la administración: los ejecutivos con acceso a todo representan un gran riesgo si sus cuentas se ven comprometidas o si involuntariamente se exponen las redes comerciales.

- **Plan de respuesta a incidentes**
  Por último, pero no menos importante, debe tenerse un proceso para manejar las brechas de seguridad si algo supera las defensas. Si bien es una buena idea subcontratar al menos parte de la seguridad a un tercero de confianza que pueda brindar respuestas a pedido, también vale la pena invertir en herramientas de borrado remoto que puedan llegar a los dispositivos móviles independientemente de su ubicación física.

Los datos externos

Los datos internos son información generada desde dentro del negocio, que cubre áreas como operaciones, mantenimiento, personal y finanzas. Los datos externos provienen del mercado, incluidos clientes y competidores. Son cosas como estadísticas de encuestas, cuestionarios, investigaciones y comentarios de los clientes.
Los analistas comerciales consideran que los datos generados internamente son más valiosos porque al poder manipularse sus variables, pueden beneficiar a las empresas que necesitan mejorar la eficiencia, la productividad y a las empresas que no logran generar ganancias.
Por el contrario, los datos externos están fuera del control de una organización, como las tendencias económicas y las regulaciones gubernamentales dentro de una industria.
Entonces, los datos internos ayudan a administrar el negocio y optimizar las operaciones. Los datos externos ayudan a comprender mejor la base de clientes y el panorama competitivo. Es necesaria una visión clara de ambas fuentes para tener una inteligencia empresarial.

Los datos externos son cualquier dato generado desde fuera de una organización. Puede provenir de una variedad de fuentes, y las iniciativas de datos abiertos (Open data) ahora son abundantes, lo que pone a disposición una gran cantidad de datos externos para su análisis.

Datos de redes sociales: Las redes sociales son una de las principales fuentes de datos externos actuales, sirven tanto para entender mejor el mercado de nuestro negocio, la percepción de los usuarios y saber qué opinión tiene el público sobre nuestra empresa, nuestros productos, algún lanzamiento reciente, etc.

Datos demográficos: El comportamiento del cliente varía según su ubicación. Esto puede deberse a variables de lo más diversas, como la edad de los compradores de cada región, su poder adquisitivo, o alguna relación más compleja de distintos datos demográficos. Poder comparar estos comportamientos nos ofrece otra perspectiva con la que tomar decisiones.
Hay sitios que disponen de una API que nos ofrece datos semiestructurados que podemos automatizar y tratar de manera eficiente.
(Una API es un paquete armado que puede ser utilizado por otros programas, se aplica a funciones, herramientas que pueden utilizar los programadores y que les ahorran trabajo.)

Datos meteorológicos: Por ejemplo, se podría comparar la temperatura y las precipitaciones que hay en el momento de la compra de un producto de nuestra empresa contra el histórico de pedidos, y ver si el clima influye en nuestro volumen de pedidos, qué tipos de clientes siguen haciendo pedido, de qué artículos y así obtener patrones.

Calendario laboral y festivo: Muestra las tendencias asociadas a patrones de comportamiento, traslados y consumo relacionadas con vacaciones y festividades.

Valores en bolsa, ya sea porque nuestra empresa cotiza en bolsa, o porque dependemos o usamos productos que sí lo hacen. Predecir valores futuros es una tarea compleja, pero un análisis que relaciones nuestros propios datos, con los cambios de bolsa, puede ampliar el panorama.

Iniciativas Open Data, se podría resumir como una iniciativa para hacer accesibles al público general los datos de distintos gobiernos a distintos niveles (municipalidades, ministerios, gobiernos, institutos oficiales de estadística o instituciones como el Banco Mundial o la Comisión Europea).

Datos de otras empresas
Hasta ahora estuvimos hablando de datos de acceso libre, pero también existen los datos recolectados por otras empresas, y que pueden reforzar los nuestros:

- **Empresas de transporte** (ferroviarias, aerolíneas, empresas especializadas en el sector hotelero, alquiler y compra de inmuebles)

pueden indicar el potencial adquisitivo de nuestros clientes, o el tipo de perfil de los mismos (en función de si viajan mucho o no, por ejemplo).

La mayoría de las **empresas telefónicas** son proveedoras de internet, ya sea mediante wifi o datos móviles, y recogen, quiero creer que de forma anónima para cumplir con las normas de protección de datos, los datos sobre el consumo de sus clientes. Esto nos da el potencial para ayudarnos a entender el perfil de los usuarios y, por tanto, de los habitantes de una determinada zona.

- **Datos especializados** nos pueden servir para encontrar información muy concreta, como pueden ser precios y cotizaciones de productos financieros, tipos de cambio monedas y precios de materias primas.
- **Bloomberg** es una compañía estadounidense que ofrece software financiero, datos y noticias en tiempo real sobre las economías de todo el mundo.
- **Nielsen Holdings Inc.** es una empresa estadounidense de medición de información, datos y mercado.

2.5 Roles en ciencia de datos

Tareas del profesional en Big Data

Un especialista en Big Data es un profesional que cuenta con amplios conocimientos en una serie de tareas involucradas en el ciclo de vida de la gestión de los datos tales como:

- ✓ identificar diversos orígenes de información,
- ✓ almacenar y extraer grandes volúmenes de datos,
- ✓ diseñar la arquitectura del ecosistema empresarial donde se procesa y consumirá los datos para su exploración,
- ✓ modelado, análisis, visualización y monitorización en tiempo real.

Dependiendo de sus funciones, un especialista en Big Data debe poseer habilidades empresariales, técnicas y analíticas para obtener el mayor provecho de la información.

Dado que el uso de plataformas de Big Data aumenta cada vez más para dar paso a la transformación digital, es común que las empresas desarrollen sus propios sistemas con componentes *legacy*, en la nube o en ambos, por lo que los expertos de Big Data deben tener dominio en diferentes lenguajes de programación, aplicaciones tecnológicas, pero además de herramientas en entornos cloud.

Las empresas con proyectos de Big Data pueden necesitar un equipo de especialistas en para manejar el flujo de trabajo de un proyecto, por ello existen diferentes perfiles con diferentes funciones y responsabilidades específicas, que podrían variar según los requisitos de cada empresa.

Conoceremos algunos de estos perfiles a lo largo del curso.

Interdisciplina

El equipo de Data Science debe tener conocimientos y habilidades en tres áreas importantes:

1. Matemáticas y estadística (estadística descriptiva - medidas de tendencia central)

2. Ciencias computacionales (programar)
3. Conocimiento del dominio de la industria (entretenimiento, educación, finanzas, etc.)

Algunas actividades se solapan con las de otros roles. La diferencia reside en la profundidad de estudio de cada área, que dependerá de la etapa del proceso a la que nos dediquemos. También debemos tener en mente el tamaño de la organización, la envergadura del proyecto, y otras variables que afectan a la conformación del equipo de trabajo.

Data Analyst

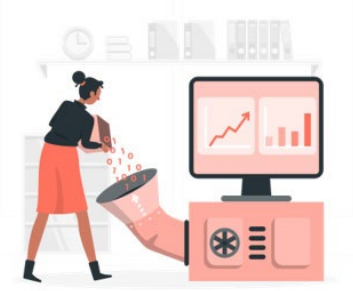
Su función en analizar el presente de una organización. Ejecuta análisis de datos para generar informes en dashboards con tablas y gráficas que ayuden aotras personas de la organización a tomar mejores decisiones o saber si alguna estrategia está funcionando.

Su principal misión es extraer datos recolectados y analizarlos. Para ello su día a día tiene estas actividades:

- ✓ Colaborar con managers y otras personas de la organización para identificar necesidades de información.
- ✓ Extraer datos de fuentes con SQL o Python.
- ✓ Limpiar y organizar los datos para su análisis.
- ✓ Analizar los datos para identificar patrones y tendencias que se puedan convertir en información accionable.
- ✓ Comunicar los hallazgos en tableros con visualizaciones fáciles de entender para la toma de decisiones y generación de estrategias.

A diferencia de una Data Scientist, una Data Analyst no suele utilizar machine learning ni colabora con ingeniería para incorporar datos a los productos, sino que se enfoca en analizar el presente de la organización. Responde los requisitos de información de colaboradores buscando datos en las bases de datos de la organización, analizándolos y reportándolos en gráficas y tablas.

¿Qué debe saber un Data Analyst?



Una Data Analyst debe conocer fundamentalmente manejo de bases de datos SQL para consulta de datos y hojas de cálculo con Excel.

Dentro de las matemáticas que debe conocer encontramos la estadística y la probabilidad.

De igual manera conoce de programación con Python utilizando librerías como Pandas, Matplotlib y Seaborn para análisis y visualización de datos.

También utiliza herramientas avanzadas de visualización y análisis de datos como Microsoft Power BI y Tableau. Estas herramientas permiten crear dashboards para consulta de información por cualquiera que forme parte de la organización en la que trabaje.

¿Cómo empezar a aprender análisis de datos?
Lo primero que necesitarás aprender es:

- ✓ Cómo utilizan los datos las organizaciones con Business Intelligence.
- ✓ Consultar bases de datos con SQL.
- ✓ Uso de herramientas para análisis de datos como Excel, Microsoft Power BI y Tableau.
- ✓ Estadística.

Data Engineer

Crean y mantienen una estructura de software que permita el procesamiento de grandes cantidades de datos que vienen de distintas fuentes de la organización y que serán usados exclusivamente para analítica de datos. Este proceso se conoce como ETL por sus siglas en inglés de extracción, transformación y carga.
El rol de Data Engineer trabaja para que los demás roles en un equipo de Data Science tengan datos para analizar.
Se preocupan en crear flujos ETL (Extracción, Transformación y Carga de datos) para que analistas y científicas de datos puedan recuperar fácilmente los datos desde bases de datos especializadas para análisis.
Su día a día consiste en las siguientes actividades:

- ✓ Desarrollar y mantener bases de datos y data pipelines de ETL que manejan gran volumen de datos brutos.
- ✓ Extraer datos de diferentes fuentes como bases de datos estructuradas y no estructuradas, API y archivos.
- ✓ Preparar los datos para que sean usados para análisis.
- ✓ Almacenar los datos en data warehouse.
- ✓ Crear automatizaciones para ejecutar periódicamente esos procesos.

¿Qué debe saber un Data Engineer?
Para desempeñar el rol de Data Engineer necesitarás principalmente saber programación con Python, bases sólidas de ingeniería de software y de uso defuentes de datos estructurados (SQL) y no estructurados (NoSQL).
Para crear los procesos de ETL usarás herramientas como como Apache Spark para la manipulación y transformación de forma paralela de Big Data,y Apache Airflow para automatizar estos procesos.
También utilizarás servicios cloud como AWS, Microsoft Azure y Google Cloud Platform para realizar todo esto dentro de la nube.
¿Cómo empezar a aprender ingeniería de datos?
Las primero que necesitarás aprender es:

- ✓ Programación con Python y bases sólidas de ingeniería de software.
- ✓ Automatización y scripting.
- ✓ Uso de librerías de Python para manipulación y análisis de datos y Apache Spark.
- ✓ Conocimientos en bases de datos SQL y NoSQL.

Data Scientist

Se encarga de tomar datos de las fuentes de información de la organización, de limpiarlos, procesarlos, analizarlos, utilizar modelos de inteligencia artificial para resolver preguntas interesantes que surjan en su organización para toma de decisiones.

Se encargan de entender al negocio y sus datos para agregar valor a la organización con toma de decisiones basadas en datos e incorporar datos a los productos de software.
Para ello su día a día contiene actividades como las siguientes:

- ✓ Obtener, limpiar y procesar datos estructurados y no estructurados de distintas fuentes.
- ✓ Diseñar y utilizar modelos de machine learning para generar predicciones sobre los datos.
- ✓ Desarrollar herramientas para monitorear la precisión de los datos.
- ✓ Automatizar procesos para recolectar y transformar datos que utilicen.
- ✓ Crear reportes en tableros con visualizaciones de información valiosa.
- ✓ Ayudar a incorporar datos a los productos de la mano con el equipo de ingeniería.

¿Qué debe saber un Data Scientist?
Sus habilidades contemplan herramientas como Python, como uno de los principales lenguajes de programación, y sus librerías como Pandas, NumPy y Matplotlib para análisis de datos y creación de gráficas que ayudan a contar historias para que en la organización puedan tomar decisiones basadas en datos.
Adicional a esto conocen el manejo de bases de datos estructurados (SQL) con herramientas como PostgreSQL y datos no estructurados (NoSQL) con herramientas como MongoDB o Apache Cassandra. Esto les sirve para extraer datos de las fuentes de información de la organización.
Por último, tienen un conocimiento de matemáticas y estadística aplicadas a ciencia de datos y de uso de modelos de machine learning y deep learning para generar predicciones sobre la información que analizan.
¿Cómo empezar a aprender Data Science?
Las cuatro primeras cosas que necesitarás aprender son:

1. Cómo utilizan los datos las organizaciones.
2. Lo esencial de programación con Python y sus librerías para análisis de datos.
3. Uso de herramientas para análisis de datos como Jupyter Notebooks.
4. Estadística y probabilidad aplicada a data science.

Machine Learning Engineer
Funciona más dentro de la capa de inteligencia artificial de una organización. Su tarea es escalar y robustecer modelos de inteligencia artificial para funcionar en sistemas de producción de software, que en ocasiones han sido creados por Data Scientists. Este rol se asocia mucho más que otros a conocimientos y buenas prácticas de la ingeniería de software.
Funciona dentro de equipos que construyen productos fuertemente basados en inteligencia artificial. Seguramente has experimentado este tipo de productos cuando, utilizando plataformas como Netflix, recibes recomendaciones con base en series y películas que has visto antes. Esto es una predicción de machine learning funcionando en un producto de software.
Para que esto sea posible, una Machine Learning Engineer tiene como tarea escalar y robustecer modelos de inteligencia artificial para funcionaren sistemas de producción de software.
Estas son las actividades que encontramos en su día a día:

1. Generar una evaluación extensiva de métricas de modelos de machine learning.
2. Diseñar y construir sistemas de machine learning.
3. Crear y ejecutar pruebas A/B de los modelos de machine learning.
4. Monitorear el desempeño y funcionalidad de los sistemas de machine learning.
5. Colaborar directamente con Data Scientists y otras áreas de ingeniería de software para asegurar la funcionalidad del producto final.

¿Qué debe saber un Machine Learning Engineer?

Este rol basa sus habilidades en bases fuertes de ingeniería de software. Puedes utilizar varios lenguajes de programación, pero el más utilizado es Python.

Es clave entender de estadística, probabilidad, cálculo y álgebra lineal aplicadas a ciencia de datos e inteligencia artificial.

Dentro del ecosistema de Python para inteligencia artificial utilizarás frameworks y librerías como scikit-learn, TensorFlow, Keras, PyTorch y NLTK. Para poder emplear estos frameworks será muy importante que conozcas cómo funcionan y cómo aplicar los diferentes tipos de algoritmos de inteligencia artificial.

Además, conocer herramientas cloud como AWS, Microsoft Azure y Google Cloud Platform, te dará los súper poderes para poner a funcionar tus modelos en producción.

¿Cómo empezar a aprender Machine Learning Engineering?

Las cuatro primeras cosas que necesitarás aprender son:

1. Programación con Python y bases sólidas de ingeniería de software.
2. Uso de librerías de Python para manipulación, análisis y visualización de datos.
3. Matemáticas aplicadas a data science e inteligencia artificial.
4. Aplicación de modelos de machine learning con scikit-learn.

En proceso

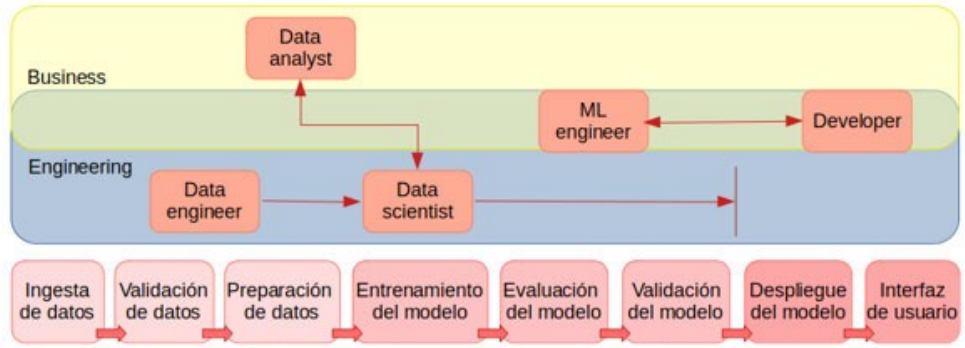
Veamos un ejemplo del proceso de Data Science para poner en producción un producto de IA con Machine Learning.

En la primera etapa de ingesta, validación, propagación de datos (recolección, limpieza, transferencia y almacenamiento en BBDD especializadas), interviene en Data Engineer.

Junto al Data Scientist, son las primeras personas en crear los modelos de Machine Learning, los primeros en interactuar con los datos. Entrenan, evalúan y validan los datos.

En la parte de negocio, más involucrado con el negocio, esta el Data Analyst, que toma los datos que prepararon los engineer y los analizan para encontrar información de valor, insights que las personas del negocio puedan usar para generar estrategias, acciones.

El Machine Learning Engineer trabaja de la mano con Data Scientist porque también interviene en la parte de evaluación y validación del modelo para robustecerlo, desplegarlo y ponerlo en producción. Adicionalmente, trabajan con desarrolladores e ingenieros de software para crear una a aplicación que pueda interactuar y ser utilizado por usuarios por medio de una interfaz.



Procesos y roles en el flujo de trabajo de Data Science

INVESTIGA Y REFLEXIONA
Estudia acerca de los roles de los profesionales en big data y responde ¿en cual te gustaría desarrollarte y porque?

2.6 Uso ético del Big Data

Los principios de privacidad global del GDMA

Las nuevas tecnologías y el uso de datos personales brindan a la humanidad la oportunidad de vivir mejor, consumir mejor y ser más sostenible. Los datos tienen un papel cada vez mayor en esta búsqueda de negocios, innovación y crecimiento económico. Los beneficios de los datos para la sociedad y la economía solo pueden lograrse a través de su uso ético y la generación de confianza entre individuos y organizaciones. Las reglas de privacidad y protección de datos contribuyen a la creación de confianza, al mismo tiempo que proporcionan un marco para los flujos de información libres y responsables en todo el mundo.

La GDMA es una organización que representa, apoya y une a las asociaciones de marketing de todo el mundo que se centran en el marketing basado en datos. Los Principios de privacidad global de GDMA establecen un marco mundial para la comunicación con el cliente que debe sustentar todos los enfoques legales y comerciales. Están diseñados como un instrumento de buenas prácticas y pretenden servir como guía para la autorregulación y la legislación.

Los Principios de Privacidad Global de GDMA son compromisos a los que aspiran organizaciones, gobiernos y personas para cultivar un ecosistema comercial confiable y exitoso a través del servicio a cada individuo con equidad, transparencia y respeto por la privacidad. El principio rector de respetar y valorar la privacidad genera confianza en el corazón de la comunicación con el cliente como un intercambio de valor entre una organización que busca prosperar y un individuo que busca beneficiarse. Estos principios garantizan que las organizaciones de todo el mundo pongan al individuo en el centro de todo lo que

hacen, de modo que se pueda confiar en las organizaciones, respetarlas y, en última instancia, sostenerlas en todos los países.

Lectura: GDMA-Global-Principles-FullText-ES-Admia-RGB.pdf