BIG DATA

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

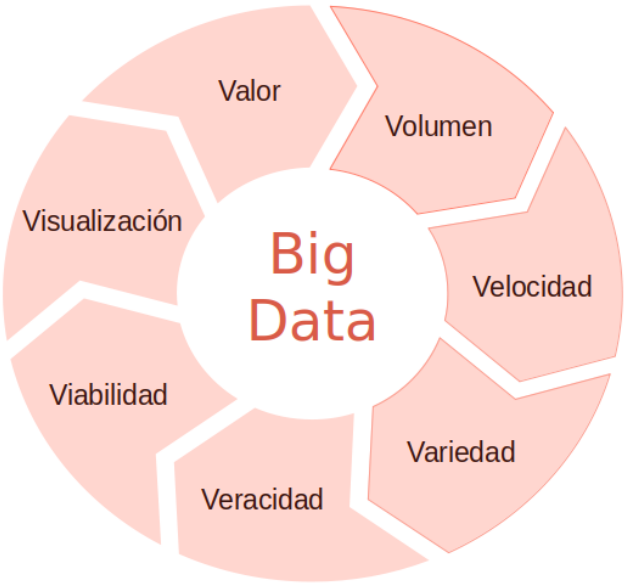
La cantidad de datos es tan grande, que las aplicaciones de software de procesamiento de datos que tradicionalmente se venían usando no son capaces de capturar, tratar y ponerlos en valor en un tiempo razonable.

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.

VIRALIZACIÓN: ¿LA 8° V?

Surge a partir de nuevos interrogantes:

- ¿Qué tan rápido se comparte o se dispersa la información?
- ¿Obtengo alguna ventaja extra al publicar la información en redes sociales?

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.

¿ESTO ES A O B?

Para responder a esta pregunta, el Data Scientist utiliza Algoritmos de clasificación. Pueden ser de clase 2, si sólo hay dos respuestas posibles (¿A o B?), o multiclase, en el caso de que haya más de dos respuestas posibles.

Permite identificar a qué categoría pertenece cierta información. Un ejemplo de estos algoritmos son los árboles de decisión. Sirve para responder preguntas como:

- ¿Esta máquina va a fallar en los próximos días?
- ¿Qué atrae más clientes, un cupón regalo o un porcentaje de descuento?
- ¿Este tweet es positivo?
- ¿Qué servicio elegirá este cliente? ¿A, B o C?

¿ES ESTO NORMAL O ACÁ PASA ALGO RARO?

Para responder a esta pregunta, se usa una familia de algoritmos que se llama detección de anomalías. Estos algoritmos identifican y predicen comportamientos que se salen de lo normal.

¿CUÁNTO O CUÁNTOS?

En este caso, podemos usar Machine Learning para predecir la respuesta mediante la familia de algoritmos de regresión. Estos algoritmos permiten predecir el valor numérico que tendrá una variable, basándose en comportamientos anteriores.

Aquí vemos un ejemplo de cómo una empresa de alquiler de autos aplicaría un algoritmo de regresión para estimar la demanda de servicios.

- **OBTENER LOS DATOS**: Variaciones de demanda. El alquiler de autos podría tener picos en ciertas horas del día, feriados, vacaciones, climas, etc.
- **PREPARAR LOS DATOS**: Depurar los datos, combinar los sets de datos y prepararlos para su análisis.
- **FORMAR EL MODELO**: Introducir la información en la máquina para enseñarle qué esperar.
- **EVALUAR EL MODELO**: Testear la habilidad del modelo para predecir los datos originales, y evaluar los aciertos.
- **PRECEDIR LA DEMANDA FUTURA**: Usar el modelo para pronosticar futuros picos y caídas en la demanda

Puede responder a preguntas como ¿Cuál será el volumen de ventas de este trimestre?

¿CÓMO ESTÁ ORGANIZADO ESTO?

Las dos últimas preguntas son un poco más complejas. A veces, necesitamos comprender la estructura de los datos, saber cómo podrían organizarse. Una de las formas más habituales para sacar a la luz la estructura de los datos es agrupándolos en conjuntos de elementos que son similares. Por ejemplo, los clientes de TV por cable pueden agruparse según el tipo de películas que les gustan. También se pueden agrupar según criterios socio-económicos (edad, sexo, nivel de estudios, situación laboral etc.). Por ello, cuando aplicamos técnicas de clustering, no hay una única respuesta correcta, sino varias que nos pueden aportar más o menos información de valor. Son muy útiles para realizar segmentaciones de clientes, predecir sus gustos o determinar un precio de mercado.

Permiten responder preguntas como éstas:

- ¿Qué modelos de impresora tienen la misma avería?
- ¿Qué libros debemos recomendar a este cliente?
- ¿Qué oferta personalizada debemos hacer a este cliente?

Si comprendemos cómo se estructuran los datos, podremos comprender y predecir comportamientos y eventos futuros.

¿Y AHORA QUÉ CONVIENE HACER?

Para responder a esta última pregunta se usa la familia de algoritmos de aprendizaje por refuerzo. Se basan en los estudios sobre cómo fomentar el aprendizaje en humanos y ratas basándose en recompensas y castigos. El algoritmo aprende observando el mundo que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error.

Estos algoritmos nos dan respuesta a preguntas que “se puede plantear” un robot o una máquina, como:

Soy un coche sin conductor y esto es un semáforo en amarillo: ¿acelero o me paro?

PROCESOS

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)

BIG DATA Y DATA SCIENCE

Big Data es la materia prima que usamos en Data Science.

Data Science es el proceso para analizar una cantidad masiva de datos en constante crecimiento, generar predicciones para la toma de decisiones y crear productos con datos. Es el proceso de descubrir información valiosa a partir de los datos, es la tecnología que usamos para entenderlos.

Los datos son lo más valioso que podemos encontrar actualmente en las empresas.

Estos datos son los que se generan cuando utilizamos alguna plataforma en línea, cuando se hace investigación, o simplemente cuando se hacen registros de algo y los almacenamos en archivos digitales.

Los datos son vitales para las empresas de hoy. Con ellos tomamos decisiones y creamos mejores productos.

LOS OBJETIVOS DE DATA SCIENCE

- 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?

La ciencia de datos involucra un proceso donde extraemos datos de diversas fuentes, los manipulamos, transformamos, visualizamos y eventualmente los usamos en modelos de Machine Learning para generar predicciones o clasificaciones. Ese tipo de modelos son parte de la inteligencia artificial.

Los pasos a seguir son los siguientes:

- Obtener los datos (mediciones directas, encuestas, internet)
- Transformar y limpiar los datos (incompletos o formato incorrecto)
- Explorar, analizar y visualizar los datos (patrones, tendencias, insights para presentar en visualizaciones o reportes amigables)
- Usar modelos de Machine Learning (IA): predecir información
- 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.

El proceso de Data Science dependerá de la empresa o proyecto en el que estemos trabajando, pero el método es siempre el mismo:

- Hacer una pregunta.

- Obtener los datos.
- Explorar los datos.
- Analizar los datos.
- Comunicar y visualizar los resultados.

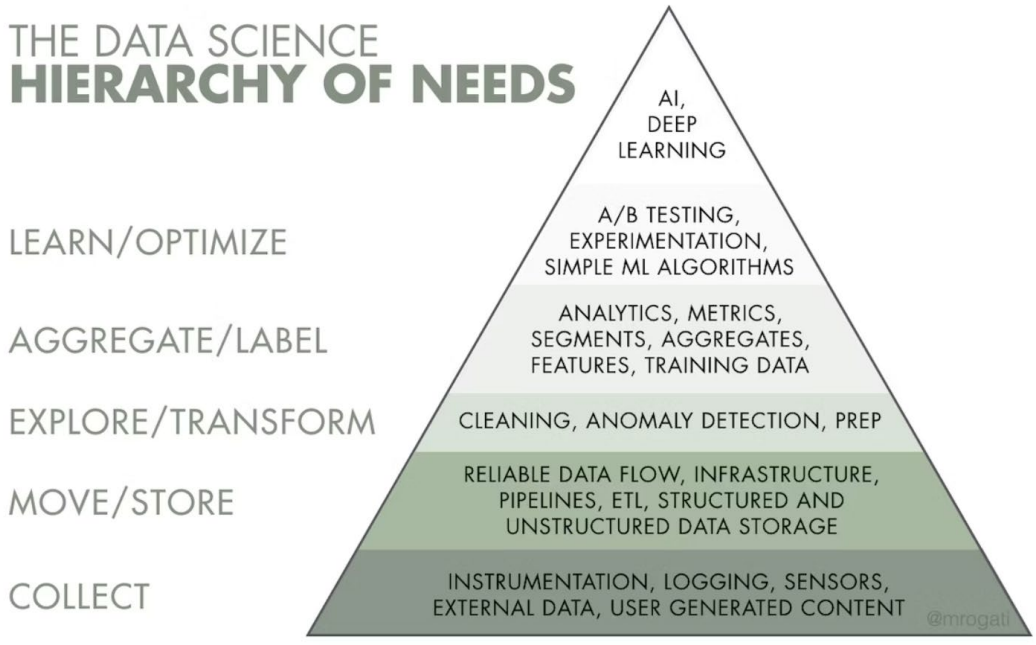
LA JERARQUÍA DE NECESIDADES EN DATA SCIENCE

Esta pirámide nos explica que no se puede implementar todo un proceso de ciencia de datos complejo, o algoritmos de Machine Learning, poner productos e Inteligencia Artificial si aún no tenemos una infraestructura, una cultura de datos para llevarlo a cabo.

Se debe partir de lo más bajo de la pirámide. Veamos cuáles son estos pasos:

- **RECOLECCION DE DATOS**: no sólo BBDD tradicionales, sino también por medio de interacciones humanas, sensores, relevamientos en sitios web, etc.
- **ALMACENAMIENTO DE DATOS**: mover los datos en una infraestructura, pipeline, ETL para obtener una base con la que podremos empezar a trabajar.
- **EXPLORACION Y LIMPIEZA**: preparar los datos para el análisis
- **ANALISIS**: preparar dashboards, gráficas y visualizaciones que ayuden a generar estrategias y tomar decisiones. Ya podemos preparar los datos para utilizarlos en un entrenamiento de Machine Learning.
- **MACHINE LEARNING**: experimentar con algoritmos simples, generar predicciones de ciertas métricas y resultados para poder ver lo que podría ocurrir a futuro.

El proceso está completo, la infraestructura y la cultura de datos están implementadas y ahora estamos listos para la Inteligencia Artificial: trabajar Deep Learning y agregar IA a nuestro producto.



ALGUNOS CONCEPTOS EN EL PROCESO DE DATA SCIENCE

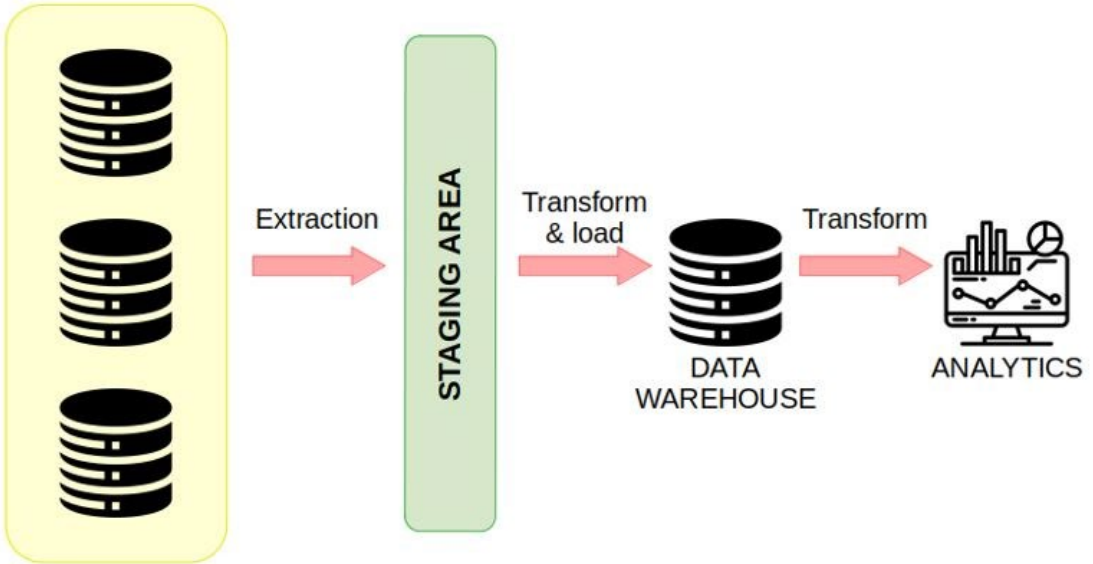
ETL (EXTRACT, TRANSFORM, LOAD)

Es el proceso de **extracción**, **transformación** y **carga de datos**. Consiste en "extraer" los datos crudos de su origen, "transformarlos" según nuestras necesidades analíticas y "cargarlos" a una BBDD orientada a procesos analíticos.

- **EXTRACCIÓN:** extraemos datos de múltiples fuentes (por ejemplo, una BBDD PostgreSQL, otra en Oracle y un archivo CSV). Es necesario conocer el formato y características de los datos, para saber la mejor manera de extraerlos. La extracción se puede hacer de dos formas:
    - **TOTAL:** en un único llamado se extrae la totalidad de datos a procesar.
    - **INCREMENTAL:** extrae los datos en pequeños lotes múltiples veces. Por ejemplo, un ETL que se ejecuta diariamente que sólo consulta los datos del día anterior.
- **TRANSFORMACIÓN:** se aplican las reglas que el negocio demande para realizar un buen proceso de analítica. Estas reglas pueden incluir procesos como:
    - Filtrar filas
    - Eliminar duplicados
    - Transformar (reemplazar) datos
    - Calcular datos nuevos (a partir de otros datos)
    - Agrupar datos (valores máximos, mínimos, promedios, conteos, etc.)
    - Unir o combinar datos de distintas fuentes
    - Pivotar las tablas
    - Dividir columnas

Estas transformaciones se realizan en la llamada "Staging Area": un repositorio temporal para procesar estos datos, que funciona por medio de tablas o archivos planos, dependiendo de la herramienta que usemos.

- **CARGA**: Es el proceso final del ETL. Los datos están transformados y listos en el área de staging. Se cargan en un una BBDD, generalmente es un data-warehouse donde conviven diferentes repositorios de datos listos para análisis de datos.



INTELIGENCIA ARTIFICIAL

Es una de las herramientas en Data Science y se trata de la inteligencia que poseen las máquinas, inspirada en la inteligencia natural que tenemos los seres humanos. Consta de algoritmos que emulan la inteligencia natural para reconocer patrones en grandes cantidades de datos. No tiene ninguna clase de conciencia. Ej: Google DeepMind, AlphaGo, sistemas de recomendación como el de YouTube, los filtros de Instagram o los asistentes de voz como Siri, Google Assistant o Alexa.

Dentro de la inteligencia artificial encontraremos tecnologías como: Machine Learning, Deep Learning, visión artificial, procesamiento de lenguaje natural, robótica, representación del conocimiento.

Un equipo de data science orientado a inteligencia artificial, además de los roles tradicionales de Data Scientist, Data Analyst y Data Engineer, comprenderá otras especializaciones, como Machine Learning Engineer, Deep Learning Engineer, NLP Scientist, Machine Learning Scientist, etc. Muchos de estos roles requieren de herramientas de data science y conocimientos más avanzados en ciencias computacionales para procesar los datos que utilizarán.

MACHINE LEARNING O APRENDIZAJE AUTOMÁTICO

Es una de las herramientas más importante de la IA. Su objetivo es que las máquinas aprendan, creando software, pasándoles algoritmos, datos y modelos para resolver problemas. Es un proceso que consta de los siguientes pasos:

- Creación del modelo
- Entrada de datos al modelo
- Ejecución de algoritmos y generación de una predicción
- Evaluación y vuelta al algoritmo

Ejemplo de aplicación: poder predecir cuándo va a haber una pérdida de clientes, buscar el mejor momento e incentivo para evitarlo.

DEEP LEARNING

Es una parte de Machine Learning donde las redes neuronales artificiales, aprenden grandes cantidades de datos. Se llama aprendizaje profundo, porque a mayor cantidad de capas interconectadas, se obtiene un aprendizaje más fino.

El campo de la inteligencia artificial es, esencialmente, cuando las máquinas pueden realizar tareas que generalmente requieren de inteligencia humana. Abarca el aprendizaje automático, donde las máquinas pueden aprender por experiencia y adquirir habilidades sin la participación de los humanos.

Deep Learning es un conjunto de algoritmos no lineales que pueden aplicarse para la modelización de datos y el reconocimiento de patrones. Cuando nos referimos a una forma no lineal estamos hablando de que las capas de las redes neuronales artificiales se apilan en una jerarquía que va desde las características de nivel de abstracción bajo hasta las de nivel de abstracción más compleja.

La forma en la que trabaja el Deep Learning es usando estas cascadas de capas con unidades de procesamiento que permiten la extracción y transformación de variables. Cada red, dentro de su jerarquía, aplica una transformación en su capa de entrada y utiliza esa información de

aprendizaje para crear un modelo estadístico de salida que itera las veces necesarias hasta lograr un nivel de precisión en el aprendizaje y respuesta aceptable.

¿CÓMO SE RELACIONAN MACHINE LEARNING Y DEEP LEARNING?

El Deep Learning se define como un área del Machine Learning que usa redes neuronales artificiales de múltiples capas. Es una herramienta muy poderosa, porque a diferencia del Machine Learning tradicional, logra aprender, de forma automática y de forma jerárquica, las features de distintos niveles de abstracción y resolver problemas difíciles de datos como: imagen, texto, video y audio.

Usando simulaciones cerebrales, se busca que:

- Haga que los algoritmos de aprendizaje sean mucho mejores y más fáciles de usar.
- Realice avances revolucionarios en el aprendizaje automático y la inteligencia artificial.

Otro punto importante es que se trata de un aprendizaje en escala. Lo que quiere decir es que a medida que construimos redes neuronales más grandes y las entrenamos con más y más datos, su rendimiento continúa aumentando. Esto generalmente es diferente a otras técnicas de aprendizaje automático que alcanzan una meseta en el rendimiento.

De manera similar a cómo aprendemos de la experiencia, el algoritmo de Deep Learning realizaría una tarea repetidamente, y cada vez que la modificara un poco se estaría buscando mejorar el resultado.

Ejemplos de aplicaciones: traducciones, asistentes virtuales, chatbots y bots de servicio, reconocimiento facial, autos sin chofer, compras y entretenimientos personalizados.

PRUEBAS A/B (SPLIT TESTING)

Son procesos que consisten en hacer experimentos mostrando dos variantes del mismo proyecto para medir y comparar cuál nos arroja mejores resultados.

Se dividen en 5 fases:

- **DEFINICIÓN DEL OBJETIVO**: Lo que pretendemos conseguir.
- **DEFINICIÓN DE HIPÓTESIS**: Qué queremos comprobar.
- **DISEÑO Y EJECUCIÓN DEL TEST**: Preparar un primer ejemplar y un segundo ejemplar, con alguna variante.
- **ANÁLISIS DE DATOS**: Si hemos elegido bien el objetivo y es cuantificable, obtendremos unos datos que deberemos analizar para llegar a nuestras conclusiones.
- **IMPLEMENTACIÓN**: Se implementa la versión que mejores resultados arroja.

STORYTELLING

Contar historias (storytelling) con nuestros datos es una manera eficaz de difundir nuestro mensaje hacia una audiencia.

Storytelling es la técnica de comunicación en la cual hilamos historias para poder comunicar un mensaje de una forma estructurada, clara y entretenida para la audiencia, de manera que pueda retenerse de una forma prolongada gracias a la captación de la atención y uso de lenguajes sencillos para explicar conceptos más complejos.

UNA BUENA HISTORIA DEBE

- Informar
- Entretener
- Inspirar

CÓMO CREAR UNA BUENA HISTORIA

- Entender la historia
- Crear la historia
- Contar la historia
- Usar la historia

OBJETIVOS

- Que se entienda
- Poner a la audiencia en el mood correcto (sembrar indicadores de cómo recibir la información)
- Que sea atractivo, en dinámica y estética
- Complementar la historia con emoción e información extra
- Que sea recordada

PILARES DE STORYTELLING

- **EL MENSAJE**: El objetivo principal siempre es el mensaje. El mensaje debe ser interpretado, no entregado. Esto ayuda a retener la atención para que la audiencia esté lista. Las historias sin mensajes no trascienden. Debemos tener bien en claro qué queremos que la audiencia entienda, sienta y haga con la historia.
- **LA CONEXIÓN**: Se trata de lograr que el mensaje importe. Conocer a la audiencia es una virtud a la hora de narrar una historia. No se trata de cómo narrar, sino de saber qué quieren y necesitan escuchar. Las mejores maneras de conectar con una audiencia son la empatía y la honestidad. Mostrarse legítimo y vulnerable abre las puertas.
- **LA EMOCIÓN**: Las emociones están en por qué y en la universalidad. Las emociones con las que nos mueven. Se alimentan a partir de partes de información. Las ideas provocan más que las propias acciones. La ideología nos hace empatizar o no con la historia.
- **LOS SENTIDOS**: Son un canal directo a nuestra memoria. Generan historias que requieren más atención e involucramiento.

LAS VÍAS PARA CREAR UNA BUENA HISTORIA

- **NARRATIVA**: Nos ofrece diversos recursos
- **PERSONA**: Punto de vista en que se narra la historia
- **TIEMPO**: Cuándo es que se narra la historia
- **NARRADOR**: Quién narra la historia
- **FORESHADOWING**: Exponer algunos eventos futuros de la historia
- **FLASHBACK**: Exponer algunos eventos pasados de la historia.
- **VISUAL**: El color, la tipografía, la iluminación, texturas y formas, montaje, encuadre, movimientos, y foco.
- **AUDIO**

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:

- Matemáticas y estadística (estadística descriptiva - medidas de tendencia central)
- Ciencias computacionales (programar)
- 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 a otras 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?

Un 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 de fuentes 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:

- Cómo utilizan los datos las organizaciones
- Lo esencial de programación con Python y sus librerías para análisis de datos
- Uso de herramientas para análisis de datos como Jupyter Notebooks
- 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 funcionar en sistemas de producción de software.

Estas son las actividades que encontramos en su día a día:

- Generar una evaluación extensiva de métricas de modelos de machine learning
- Diseñar y construir sistemas de machine learning
- Crear y ejecutar pruebas A/B de los modelos de machine learning
- Monitorear el desempeño y funcionalidad de los sistemas de machine learning
- 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:

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

EN PROCESO

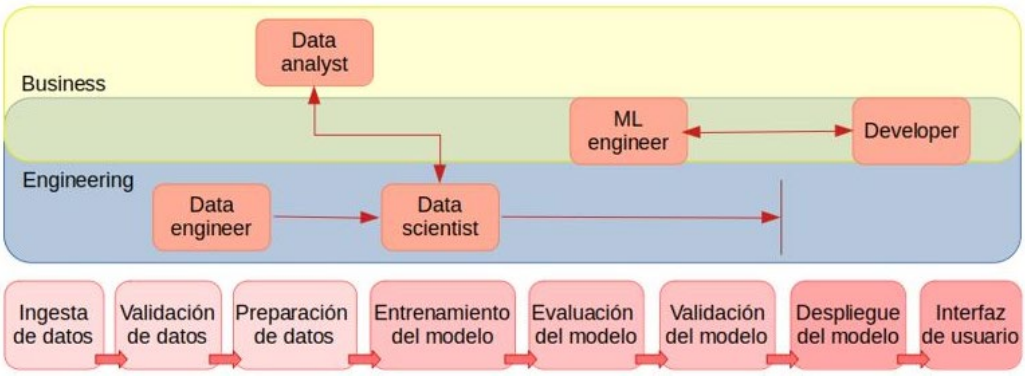
Veamos un ejemplo del proceso de Data Science para poner en producción un producto de IA con Machine Learning.

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, está 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.



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.

TIPOS Y FUENTES DE DATOS

TIPOS DE DATOS

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.

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.

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.

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.

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 el 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).