Curso

Data Engineer

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Centro de e-Learning UTN.BA

Módulo 1:

Extracción y almacenamiento de datos

Unidad 1:

Introducción

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Presentación

Antes de profundizar en aspectos técnicos de la ingeniería de datos, necesitamos contextualizar para entender de qué se trata y cuál es su importancia.

Por eso, en esta unidad temática, vamos a hablar sobre el potencial de los datos en las organizaciones para la toma de decisiones y cuál es el flujo que siguen los datos para poder generar información valuable y accionable.

Una vez definido ese flujo, podremos ver dónde interviene la ingeniería de datos y cuál es su responsabilidad fundamental, además de las funciones específicas que el/la ingeniero/a de datos puede utilizar, junto con las herramientas populares del mercado.

Esta unidad busca dar un panorama sobre la ingeniería de datos y así poder profundizar detalles técnicos en las próximas unidades temáticas.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Bloques temáticos

1. Importancia de los datos en las organizaciones
2. Ciclo de vida de los datos
3. Rol de ingeniero/a de datos
4. Día a día: Herramienta y tecnologías

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Importancia de los datos en las organizaciones

Los datos permiten **tomar decisiones** de una manera **informada** y basada en evidencias, mejor llamada **“data-driven”**. Por ejemplo, cuando vamos a salir de nuestra casa, podemos ver el pronóstico del clima para decidir qué ropa nos vamos a poner, si tenemos que llevar un abrigo adicional, o un paraguas, entre otras cosas. A partir de los datos que nos provee el pronóstico del clima, podemos tomar una decisión “data-driven”.

En un ámbito profesional, cuando una organización analiza sus datos, es capaz de descubrir patrones y tendencias y de hacer inferencias. De esa forma, podrá mejorar sus procesos o brindar mejores servicios a sus clientes. Por ejemplo, Netflix a partir de datos que recolecta de sus usuarios, personaliza el contenido que les muestra en su plataforma de acuerdo a diferentes aspectos como gustos, entorno, horario, etc. Incluso, Netflix personaliza la portada de las series y películas que le recomienda a sus usuarios a partir de dichos datos.

Las organizaciones cuentan con sistemas que generan y capturan datos. En nuestro día a día, usamos dispositivos y aplicaciones que crean y recopilan una cantidad de datos cada vez mayor.

Sumado a eso, existe una variedad de métodos y técnicas que permiten, a partir de esos datos crudos, obtener valor e información accionable para tomar decisiones.

De eso se trata la analítica de datos, abarca la recopilación, transformación y organización de datos con el fin de extraer conclusiones, hacer predicciones e impulsar la toma de decisiones fundamentadas en evidencia. Todo eso lo hace por medio de herramientas, tecnologías y procesos utilizados para resolver problemas mediante el uso de datos.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

La analítica de datos es muy importante para las organizaciones porque ayuda a obtener más visibilidad y un conocimiento más profundo de sus procesos y servicios. De esa forma pueden identificar oportunidades de mejora y optimización, lo que les llevará a reducir costes y desarrollar mejores productos y servicios centrados en el cliente.

Existen cuatro tipos principales de analíticas de datos:

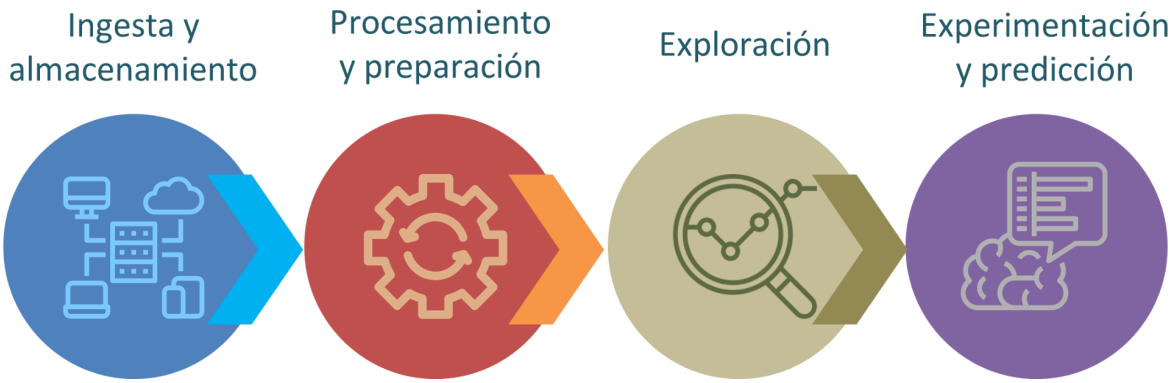
- Descriptiva: permite comprender que ocurrió o que está ocurriendo en el entorno de datos. Por lo general, se realiza por medio de visualizaciones como gráficos de línea, de torta, de barras, etc.
- Diagnóstica: permite entender por qué sucedió algo, por medio de técnicas de “data mining” y de operaciones y transformaciones sobre los datos.
- Predictiva: puede decir lo que probablemente ocurrirá en el futuro a partir de los datos históricos por medio de machine learning, modelado predictivo, entre otras cosas.
- Prescriptiva: no sólo predice lo que es probable que ocurra, sino que también sugiere una respuesta óptima a ese resultado, es decir ofrece el mejor curso de acción para atacar el problema. Se basa en análisis de grafos, redes neuronales, motores de recomendación, etc.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Ciclo de vida de los datos

En Data Analytics, o al momento de trabajar con datos, el flujo de trabajo se divide en cuatro principales etapas:

- Ingesta y almacenamiento
- Procesamiento y preparación
- Exploración
- Experimentación y predicción



En cada una de estas etapas intervienen distintos roles o profesionales:

Etapa	**Roles**
Ingesta y almacenamiento	Ing. de datos
Procesamiento y preparación	Ing. de datos
Exploración	Análisis de datos Ciencia de datos
Experimentación y predicción	Ciencia de datos

Veamos, en un alto nivel, en qué consisten cada una de estas etapas:

Ingesta y almacenamiento

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Los datos de las organizaciones pueden estar dispersos en diferentes sistemas, bases de datos y repositorios. El primer paso es recolectarlos de forma periódica y centralizarlos en un repositorio o base de datos. De eso se encarga la etapa de ingesta y almacenamiento.

La ingesta es el proceso responsable de la obtención de los datos de diferentes sistemas y orígenes para depositarlos en algún sistema de almacenamiento, que puede ser un repositorio de archivo o un sistema de base de datos. El principal objetivo es adquirir los datos y disponibilizarlos en algún lugar para permitir la limpieza y exploración de los datos.

Procesamiento y preparación

Luego que los datos hayan sido recolectados, sigue el procesamiento. El procesamiento toma los datos crudos, los limpia y los convierte en un formato más significativo. El resultado es un conjunto de datos limpio y organizado que puedes utilizar para realizar consultas y generar visualizaciones, dándole la forma y el contexto necesarios para ser interpretados. Algunas de las cosas que se realizan en el procesamiento es la eliminación de registros duplicados, nulos o erróneos, conversión de columnas de un tipo de datos a otro, creación de nuevas columnas a partir de cálculos, unión de tablas para complementar información, etc.

El objetivo es obtener datos lo suficientemente coherentes para que el equipo de análisis y ciencia pueda explorarlos y realizar distintos tipos de analítica sin ningún inconveniente.

Exploración

Una vez que los datos han sido procesados, es posible interactuar con ellos, plantear hipótesis y responderlas por medio de los datos disponibles. Por lo

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

general, en este punto entra en juego la analítica descriptiva y la diagnóstica. Por medio de técnicas estadísticas, de minería de datos y de “data wrangling”, es posible descubrir tendencias y patrones.

Experimentación y predicción

Una vez que se analizaron los datos históricos, se obtuvieron conclusiones y se descubrieron tendencias, el siguiente nivel es la analítica predictiva.

En esta etapa, entra en juego el machine learning. Con los datos disponibles, se realizan “experimentos” para ver cuál es el modelo o algoritmo correcto para generar predicciones.

Rol de Ingeniero/a de datos

La ingeniería de datos se enfoca en las primeras dos etapas del flujo de trabajo Data Analytics: la de ingesta y almacenamiento y la de procesamiento y preparación. Por ello, tiene un rol fundamental y de gran importancia ya que sientan las bases para permitir distintos tipos de analítica. Si los datos están dispersos, corruptos y son de difícil acceso, no hay mucho que preparar, explorar o con lo qué experimentar.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

La ingeniería de datos es responsable de **ingestar y almacenar datos** de diversos orígenes, de distintos formatos y de distintas estructuras, **para que sean fácilmente accesibles y listos para ser analizados y explotados.**

El/la ingeniero/a debe entregar los datos correctos, en la forma correcta, a la/s persona/s correctas, de la manera más eficiente posible. Los datos entregados deben ser datos actualizados, precisos y relevantes.

Ingesta de datos

Una de las funciones de la ingeniería de datos es la ingesta, o recolección, de datos. Como sabemos, las organizaciones generan y capturan datos por medio de sus sistemas y aplicaciones. El/la ingeniero/a de datos debe extraer datos de los diferentes sistemas y fuentes de las organizaciones.

Dichas fuentes son muy variadas, van desde fuentes de datos estructuradas hasta fuentes no estructuradas. Nos podemos encontrar con bases de datos relacionales, no relacionales, APIs, repositorios con archivos, ya sea archivos CSV, Excel, imágenes, PDFs, etc.

Por lo general, los procesos de ingesta o extracción, depositan los datos en algún repositorio central como un Data Lake o un Data Warehouse.

En la siguiente unidad temática profundizaremos sobre los diferentes orígenes de datos posibles en una organización y cómo extraer datos de los mismos.

Procesamiento y preparación

Una vez que los datos han sido recolectados y almacenados en algún repositorio central. El siguiente paso consiste en procesarlos y prepararlos para darles el formato adecuado para permitir el análisis y la ciencia de datos de una forma eficiente.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

La etapa de procesamiento también suele llamarse “Transformación”. Por un lado, está orientada a la limpieza de los datos ya que nos podemos encontrar con registros corruptos, con valores nulos o incluso repetidos. La limpieza busca eliminar o darle algún tratamiento a dichos datos con inconsistencias.

Por otro lado, en esta etapa se busca estandarizar los datos. Por ejemplo: al haber varios orígenes de datos, cada uno puede manipular las fechas con un formato específico e incluso con una zona horaria diferente. Entonces es necesario convertir esas fechas a algún formato estándar definido por la organización o por el mismo equipo de Data.

También, en el procesamiento y la transformación, se busca modelar los datos para darle una estructura entendible a los/as analistas y científicos/as. Allí es necesario aplicar lógica para crear nuevas columnas o cruzar diferentes datos.

Por último, el equipo de Ing. de datos debe entregar los datos en un formato óptimo para que el equipo de análisis y ciencia de datos pueda consumirlos y experimentar con ellos sin tanta latencia, por ejemplo.

La fase de procesamiento es muy flexible e importante para entregar datos de calidad.

Automatización y orquestación

Los desarrollos que podemos hacer como ingeniero/a de datos pueden ser denominados “Data Pipelines”. Un conjunto ordenado de procesos o rutinas que se encargan de obtener, procesar, verificar y entregar datos. Dichos pipelines deberían ejecutarse de forma automática sin la intervención humana para que el consumidor lo vea en algún horario y fecha específico o con cierta periodicidad.

Aquí viene otra responsabilidad de Data Engineering: automatizar los pipelines por medio de ciertas plataformas para que los procesos se ejecuten de forma periódica. A eso, se suma la orquestación, que consiste en lanzar la ejecución

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

de los procesos del pipeline de forma ordenada por medio de alguna plataforma que, además tome alguna acción si alguna de las fases de la cadena falla.

Día a día: Herramienta y tecnologías

En el apartado anterior, hemos conocido las responsabilidades de un/a ingeniero/a de datos. En una jornada de trabajo, podemos realizar extracciones de una base de datos relacional, obtener archivos de algún repositorio como Sharepoint, chequear los tipos de datos de las columnas y convertirlos a uno más adecuado, crear una lógica para unir diferentes tablas y crear nuevas columnas a partir de cálculos, crear un modelo de datos que unifique datos de diferentes sistemas, automatizar y orquestar un pipeline para ejecutarlo cada hora, etc.

A continuación, vamos a ver qué tecnologías o herramientas utiliza para cada una de sus tareas.

En primer lugar, las dos principales herramientas que forman parten del día a día de un/a ingeniero/a de datos son:

- Un lenguaje de programación como Python (o Scala, Java, etc.), principalmente para la manipulación de los datos. Se necesita una herramienta flexible para aplicar de negocio sobre los datos.
- SQL. Las organizaciones suelen tener muchas bases de datos relacionales y el SQL es el lenguaje para consultar y obtener los datos.

Ingesta de datos

Necesitamos una herramienta que facilite la conexión a diferentes sistemas, como bases de datos SQL, APIs, servidores SFTP, etc.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Para la extracción nos podemos encontrar con plataformas como: Airbyte, Apache Nifi, Fivetran, Azure Data Factory, etc. O bien podemos usar algún lenguaje de programación como Python, con la diferencia de que las herramientas mencionadas antes ya cuentan con conectores pre-fabricados para extraer los datos de una forma más sencilla y rápida.

Almacenamiento

Los datos deben concentrarse en algún sistema central. Para centralizar los datos crudos como archivos podemos usar tecnologías como Apache Hadoop, Amazon S3, Azure Data Lake Storage, Google Cloud Storage, Minio. Ahora bien, para permitir analítica los datos deben entregarse en alguna base de datos OLAP como Apache Hive, Amazon Redshift, Google BigQuery, Azure Synapse, Apache Pinot, Apache Druid, Apache Impala, etc.

Procesamiento de datos

Al encontrarnos con grandes volúmenes de datos, el procesamiento suele enfocarse en herramientas que usen la memoria RAM para agilizar los tiempos de ejecución. Aquí podemos utilizar tecnologías como Apache Spark, Apache Flink, Apache Beam, entre otros.

Orquestación

Por último, para automatizar y asegurar la ejecución de los procesos de forma ordenada, las tecnologías de orquestación disponibles pueden ser: Apache Airflow, Prefect, Azure Data Factory, Dagster, etc.

El universo de tecnologías es muy amplio y constantemente surgen nuevas tecnologías que se van popularizando y adoptando en el mercado. Al dominar las técnicas de ingeniería de datos, la adopción de cualquier tecnología es bastante fácil.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Conclusión

En esta unidad hemos tratado la importancia de los datos en las organizaciones. Por medio de Data Analytics, es posible generar información valuable y accionable para las organizaciones y una toma de decisiones efectiva y basada en evidencia.

En el contexto de Data Analytics, los datos tienen un ciclo de vida hasta llegar a un modelo predictivo o un análisis estadístico. La ingeniería de datos aparece en la primera parte del flujo para posibilitar el trabajo de los analistas y científicos de datos y entregar los datos de la mejor manera posible.

En esta unidad explicamos dónde entra en juego la ingeniería de datos, sus responsabilidades y tareas, además de las tecnologías que facilitan su trabajo.

En las siguientes unidades profundizaremos, desde el punto de vista técnico, las diferentes responsabilidades del/a ingeniero/a de datos.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Bibliografía utilizada y sugerida

- Data Engineering Concepts, Processes, and Tools. (2019, diciembre). *AltexSoft*. https://www.altexsoft.com/blog/datascience/what-is-data-engineering-explaining-data-pipeline-data-warehouse-and-data-engineer-role/
- *Describe data ingestion and processing - Training*. (s. f.). Microsoft Learn. https://learn.microsoft.com/en-us/training/modules/explore-concepts-of-data-analytics/2-describe-data-ingestion-process
- The State of Data Engineering 2022. (2022, 7 noviembre). *lakeFS*. https://lakefs.io/blog/the-state-of-data-engineering-2022/
- *What is Data Analytics? - Data Analytics Explained - AWS*. (s. f.). https://aws.amazon.com/what-is/data-analytics/

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning