

# **Fundamentos de Ingeniería de Datos**

Diego Saavedra

Oct 4, 2023

# Table of contents

<b>1</b>	<b>¿Por qué leer este libro?</b>	<b>3</b>
<b>2</b>	<b>¿Qué busca dar a conocer este libro?</b>	<b>4</b>
<b>3</b>	<b>¿Para quién está dirigido este libro?</b>	<b>5</b>
<b>4</b>	<b>¿Cómo puedes contribuir?</b>	<b>6</b>
<b>I</b>	<b>Información General</b>	<b>7</b>
<b>5</b>	<b>Índice de Contenidos</b>	<b>8</b>
5.1	Introducción . . . . .	8
5.2	Fundamentos de DataOps . . . . .	8
5.3	Programación y Procesamiento de Datos . . . . .	9
5.4	Evolución en la Carrera de Data Engineer . . . . .	10
5.5	Conclusiones . . . . .	10
5.6	Apéndices y Recursos Adicionales . . . . .	11
5.7	Glosario . . . . .	11
5.8	Índice de Figuras y Tablas . . . . .	11
5.9	Agradecimientos . . . . .	11
<b>II</b>	<b>Introducción</b>	<b>12</b>
<b>6</b>	<b>¿Qué es Ingeniería de Datos?</b>	<b>13</b>
6.1	Importancia en el mundo actual de la tecnología . . . . .	13
6.2	Rol de la Ingeniería de Datos en el procesamiento de información . . . . .	13
<b>7</b>	<b>¿Qué aprendimos?</b>	<b>15</b>
<b>8</b>	<b>¿Qué es un Data Engineer?</b>	<b>16</b>
8.1	Importancia en el Contexto Actual . . . . .	16
8.2	Perfil y Responsabilidades de un Data Engineer . . . . .	16
8.3	Habilidades y Competencias Requeridas . . . . .	17
8.4	Oportunidades Laborales en el Campo de la Ingeniería de Datos . . . . .	17
<b>9</b>	<b>¿Qué aprendimos?</b>	<b>19</b>

# 1 ¿Por qué leer este libro?

Bienvenido a **Fundamentos de Ingeniería de Datos**, una guía amigable y completa que te sumergirá en el emocionante mundo de la Ingeniería de Datos. Tanto si eres un principiante curioso como un profesional en crecimiento, este libro está diseñado para ti. Aquí, explorarás los pilares fundamentales, las herramientas esenciales y las prácticas más efectivas para trabajar con datos.

A lo largo de estas páginas, te embarcarás en un viaje desde los conceptos más básicos de la Ingeniería de Datos hasta el diseño y la automatización de tuberías de datos, la colaboración con diversas fuentes de datos, la implementación de estrategias de DataOps y la evolución hacia roles especializados y de liderazgo en el mundo de los datos.

Lo mejor de todo es que no solo nos quedaremos en la teoría. Este libro te brindará ejemplos y ejercicios prácticos que podrás aplicar en situaciones reales. También te ayudará a vislumbrar las emocionantes oportunidades profesionales en el campo de la Ingeniería de Datos y cómo avanzar en tu carrera.

Así que, sin importar si estás comenzando tu viaje en el universo de los datos o deseas fortalecer tus habilidades existentes, **Fundamentos de Ingeniería de Datos** te proporcionará los conocimientos y las herramientas necesarias para triunfar en este apasionante y en constante evolución campo.

## 2 ¿Qué busca dar a conocer este libro?

Nuestra misión principal con este libro es compartir contigo una comprensión sólida de los fundamentos de la Ingeniería de Datos. Queremos que descubras los conceptos esenciales, las prácticas óptimas, las herramientas vitales y las habilidades que necesitas para destacar en el campo de la Ingeniería de Datos. Además, estamos aquí para guiarte en tu desarrollo profesional, desde los primeros pasos hasta roles avanzados y de liderazgo.

### 3 ¿Para quién está dirigido este libro?

Este libro está diseñado para una amplia audiencia, desde:

- Aquellos que sueñan con convertirse en Data Engineers y buscan una introducción sólida y práctica.
- Profesionales que trabajan con datos en diferentes capacidades, como analistas o científicos de datos, que desean profundizar en la adquisición y preparación de datos.
- Data Engineers en ciernes que desean consolidar sus habilidades y aprender sobre el crecimiento de su rol.
- Individuos con experiencia en tecnología de la información y desarrollo de software que desean especializarse en Ingeniería de Datos.
- Líderes y gestores de equipos de datos que desean comprender mejor el trabajo de sus equipos y mejorar su gestión.

## 4 ¿Cómo puedes contribuir?

¡Tu contribución es valiosa! Puedes ayudar de muchas maneras:

- Aprende y aplica los conceptos en proyectos reales.
- Haz preguntas y comparte tus comentarios para enriquecer el contenido.
- Comparte tus conocimientos y recursos adicionales con la comunidad.
- Sugerencias de mejora, como ejemplos adicionales o temas que consideres relevantes.
- Aplica lo que aprendas en situaciones del mundo real y comparte tus éxitos y desafíos con la comunidad.

En resumen, este libro busca ser una herramienta interactiva y colaborativa de aprendizaje. Tu experiencia y participación enriquecerán el contenido y beneficiarán a otros apasionados por la Ingeniería de Datos. ¡Bienvenido a bordo de este emocionante viaje!

**Part I**

**Información General**

# 5 Índice de Contenidos

## 5.1 Introducción

- ¿Qué es Ingeniería de Datos?
  - Definición de Ingeniería de Datos
  - Importancia en el mundo actual de la tecnología
  - Rol de la Ingeniería de Datos en el procesamiento de información
- ¿Qué es un Data Engineer?
  - Perfil y responsabilidades de un Data Engineer
  - Habilidades y competencias requeridas
  - Oportunidades laborales en el campo de la Ingeniería de Datos
- Cómo Convertirse en Data Engineer
  - Ruta de aprendizaje y desarrollo profesional
  - Recursos recomendados para aprender Ingeniería de Datos
  - Consejos para el éxito en la carrera de Data Engineer
- Dónde Ejercer como Data Engineer
  - Sectores y empresas que demandan profesionales en Ingeniería de Datos
  - Posibilidades de trabajo remoto y a nivel internacional

## 5.2 Fundamentos de DataOps

- Tareas de Data Engineer: DataOps
  - Concepto y principios de DataOps
  - Integración de desarrollo y operaciones en la gestión de datos
  - Beneficios de la implementación de DataOps
- Herramientas del Ciclo de DataOps
  - Visión general de las herramientas esenciales
  - Automatización, control y seguimiento en DataOps
- Agile en Ingeniería de Datos
  - Metodologías Agile y Scrum aplicadas a proyectos de datos
  - Desarrollo iterativo y adaptativo en Ingeniería de Datos

## 5.3 Programación y Procesamiento de Datos

- Lenguajes de Programación e Ingeniería de Software
  - Lenguajes comunes en Ingeniería de Datos
  - Desarrollo de software orientado a datos
- Dónde y Cómo Escribir Tu Código en Ingeniería de Datos
  - Entornos de desarrollo y herramientas de programación
  - Prácticas recomendadas para el código en proyectos de datos
- Automatización y Scripting
  - Automatización de tareas repetitivas
  - Uso de scripts en la manipulación y transformación de datos
- Fuentes de Datos: SQL, NoSQL, API y Web Scraping
  - Tipos de fuentes de datos y sus características
  - Métodos para acceder y extraer datos de diferentes fuentes
- Procesamiento de Datos: Pipelines, Apache Spark y Cómputo Paralelo
  - Concepto de pipelines de datos
  - Introducción a Apache Spark y procesamiento paralelo
- Automatizar los Pipelines: Airflow
  - Uso de Apache Airflow para orquestar flujos de trabajo de datos
  - Programación y programación de tareas en Airflow
- Containers y Empaquetamiento: Docker y Kubernetes
  - Virtualización de entornos de desarrollo y producción
  - Gestión de contenedores con Docker y Kubernetes
- Manejo de Ambientes para Datos
  - Creación y gestión de ambientes de desarrollo y producción
  - Separación de ambientes y control de versiones
- Testing de Software y de Datos
  - Importancia del testing en proyectos de Ingeniería de Datos
  - Estrategias y mejores prácticas de testing de datos
- CI/CD Básico (Continuous Integration/Continuous Deployment)
  - Implementación de CI/CD en proyectos de Ingeniería de Datos
  - Automatización de pruebas y despliegue de datos
- Servidores y Computación en la Nube para Data

- Uso de servicios en la nube para almacenamiento y procesamiento de datos
- Escalabilidad y flexibilidad en la nube
- Reentrenamiento y Control de Salud de Servicios
  - Mantenimiento y mejora continua de sistemas de datos
  - Monitoreo y diagnóstico de problemas
- Medición de Indicadores y Seguimiento a Proyectos
  - Establecimiento de métricas y KPIs en proyectos de datos
  - Seguimiento del rendimiento y evaluación de resultados

## 5.4 Evolución en la Carrera de Data Engineer

- Ejerciendo como Data Engineer
  - Experiencias y desafíos en el trabajo diario de un Data Engineer
  - Colaboración con otros profesionales y equipos
- Buscando Oportunidades como Data Engineer
  - Estrategias para encontrar empleo y proyectos freelance
  - Construcción de un portafolio atractivo
- Evolución en el Rol: Ganando Seniority como Data Engineer
  - Crecimiento y avance en la carrera de Ingeniería de Datos
  - Desarrollo de habilidades de liderazgo y mentoría
- Evolución en el Rol: Manager, Architect, Pivot
  - Posibles direcciones en la carrera de Ingeniería de Datos
  - Transición a roles de gestión, arquitectura o cambio de especialización
- Trabajando en Equipo como Data Engineer
  - Colaboración efectiva en equipos multidisciplinarios
  - Comunicación y trabajo conjunto con otros profesionales

## 5.5 Conclusiones

- Resumen de los conceptos clave en Ingeniería de Datos
- Perspectivas futuras y tendencias en el campo

## **5.6 Apéndices y Recursos Adicionales**

- Recursos Adicionales
  - Bibliografía recomendada
  - Enlaces a cursos y herramientas útiles

## **5.7 Glosario**

- Definiciones de términos técnicos y conceptos relevantes

## **5.8 Índice de Figuras y Tablas**

- Índice de Figuras y Tablas
  - Referencias a todas las ilustraciones y tablas del libro

## **5.9 Agradecimientos**

- Reconocimiento a personas y organizaciones que contribuyeron al libro
- Especificaciones técnicas y recursos necesarios para seguir los ejercicios prácticos

## **Part II**

# **Introducción**

## 6 ¿Qué es Ingeniería de Datos?

La Ingeniería de Datos es el proceso de adquisición, almacenamiento, transformación y análisis de grandes volúmenes de datos para obtener información valiosa y tomar decisiones informadas. En palabras sencillas, se trata de convertir datos crudos en conocimiento útil.

### 6.1 Importancia en el mundo actual de la tecnología

La Ingeniería de Datos desempeña un papel fundamental en la era digital actual. Con el crecimiento explosivo de los datos en línea, desde redes sociales hasta sensores de dispositivos inteligentes, su gestión adecuada se ha vuelto esencial. Permite a las empresas y organizaciones tomar decisiones estratégicas basadas en datos precisos y en tiempo real.

### 6.2 Rol de la Ingeniería de Datos en el procesamiento de información

Imagina que eres el gerente de una cadena de tiendas minoristas y deseas conocer las tendencias de compra de tus clientes para optimizar el inventario. Aquí es donde entra en juego la Ingeniería de Datos. Esta disciplina te ayudaría a:

1. Adquirir datos de ventas, preferencias y comportamientos de clientes de todas tus tiendas en línea y físicas.
2. Almacenar estos datos de manera eficiente en una base de datos diseñada para grandes volúmenes de información.
3. Transformar los datos en informes y gráficos comprensibles para identificar patrones y tendencias de compra.
4. Analizar los resultados para tomar decisiones informadas, como ajustar el inventario o lanzar campañas de marketing específicas.

#### Tip

Para profundizar en la comprensión de la Ingeniería de Datos, puedes explorar los conceptos de “ETL” (Extract, Transform, Load) que se utilizan comúnmente en este campo. ETL se refiere al proceso de extraer datos de diversas fuentes, transformarlos en el formato deseado y cargarlos en un almacén de datos para su análisis.

## **i** Actividad Práctica: Exploración de Datos

1. **Selecciona una Fuente de Datos:** Elige una fuente de datos accesible, como un conjunto de datos en línea o una hoja de cálculo con datos que te interesen (por ejemplo, registros de ventas, datos climáticos o listas de reproducción de música).
2. **Explora los Datos:** Abre los datos y comienza a explorarlos. Observa las columnas, las filas y los tipos de datos presentes. ¿Qué información contiene?
3. **Formula Preguntas:** Basándote en la exploración inicial, formula al menos tres preguntas que te gustaría responder utilizando estos datos. Por ejemplo, si estás explorando datos de ventas, podrías preguntarte: “¿Cuál fue el mes con las ventas más altas?” o “¿Hay una tendencia estacional en las ventas?”.
4. **Visualización de Datos:** Utiliza una herramienta de visualización de datos, como una hoja de cálculo o una herramienta en línea, para crear gráficos simples que te ayuden a responder tus preguntas. Por ejemplo, crea un gráfico de barras para mostrar las ventas mensuales.
5. **Conclusiones:** Después de explorar y visualizar los datos, reflexiona sobre las respuestas a tus preguntas. ¿Qué aprendiste de los datos? ¿Puedes identificar patrones o tendencias?

Esta actividad práctica te permitirá experimentar con datos reales y aplicar los conceptos básicos de la Ingeniería de Datos, como la exploración de datos y la formulación de preguntas. Aunque no estás realizando un proceso ETL completo, esta actividad es un primer paso para comprender cómo se utilizan los datos en la práctica.

## 7 ¿Qué aprendimos?

En esta sección, hemos explorado los fundamentos de la Ingeniería de Datos, comprendiendo su definición, su importancia en el mundo actual de la tecnología y su papel en el procesamiento de información. Aprendimos que la Ingeniería de Datos implica la adquisición, almacenamiento, transformación y análisis de datos para convertirlos en conocimiento útil. Además, exploramos el proceso ETL como una parte esencial de esta disciplina y lo aplicamos en una actividad práctica para comprender mejor cómo funciona en la vida real. ¡Ahora estás listo para avanzar y explorar más aspectos emocionantes de la Ingeniería de Datos!

## 8 ¿Qué es un Data Engineer?

Un Data Engineer es un profesional especializado en la gestión y transformación de datos, desempeñando un papel crucial en el ciclo de vida de los datos. Este experto es responsable de adquirir, almacenar, procesar y entregar datos de alta calidad para su análisis y uso en organizaciones y proyectos diversos.

### 8.1 Importancia en el Contexto Actual

En el mundo actual, donde los datos son la columna vertebral de la toma de decisiones, los Data Engineers son esenciales para garantizar que las organizaciones tengan acceso a datos precisos y oportunos. Su trabajo facilita la toma de decisiones basada en datos, la automatización de procesos y la creación de sistemas inteligentes.

### 8.2 Perfil y Responsabilidades de un Data Engineer

Los Data Engineers tienen varias responsabilidades clave, que incluyen:

- **Adquisición de Datos:** Recopilar datos de diversas fuentes, como bases de datos, registros de servidores, fuentes externas y más.
- **Almacenamiento de Datos:** Diseñar y mantener sistemas de almacenamiento de datos eficientes, como bases de datos relacionales, almacenes de datos o sistemas de archivos distribuidos.
- **Transformación de Datos:** Limpiar y transformar datos para que sean coherentes y adecuados para su análisis. Esto puede incluir la normalización de datos, la eliminación de duplicados y la agregación.
- **Creación de Pipelines de Datos:** Desarrollar y gestionar pipelines de datos que muevan y procesen datos de manera eficiente a través de diferentes etapas.
- **Automatización:** Automatizar tareas repetitivas para garantizar la consistencia y la eficiencia en el manejo de datos.

## 8.3 Habilidades y Competencias Requeridas

Para tener éxito como Data Engineer, se requieren las siguientes habilidades y competencias:

- **Programación:** Conocimiento sólido de lenguajes de programación como Python, Java o Scala para desarrollar scripts y aplicaciones que manipulen datos.
- **Bases de Datos:** Experiencia en bases de datos relacionales (como SQL) y no relacionales (como MongoDB) para gestionar y consultar datos.
- **Herramientas de Big Data:** Familiaridad con herramientas de procesamiento de big data como Apache Hadoop y Apache Spark.
- **ETL (Extract, Transform, Load):** Capacidad para diseñar y gestionar procesos ETL que muevan datos de una fuente a otra y los transformen según sea necesario.
- **Conocimiento de Cloud:** Familiaridad con plataformas de nube como AWS, Azure o Google Cloud para almacenamiento y procesamiento de datos en la nube.
- **Comunicación:** Habilidad para colaborar con otros equipos y comunicar de manera efectiva los requisitos y resultados de los proyectos de datos.

## 8.4 Oportunidades Laborales en el Campo de la Ingeniería de Datos

El campo de la Ingeniería de Datos ofrece numerosas oportunidades laborales, incluyendo:

- **Data Engineer:** El rol principal de gestión y transformación de datos.
- **Data Architect:** Diseñador de la arquitectura de sistemas de datos.
- **Data Analyst:** Profesional que analiza datos para extraer información valiosa.
- **Data Scientist:** Científico de datos que utiliza análisis avanzados para tomar decisiones.
- **Data Operations Engineer:** Responsable de mantener la disponibilidad y el rendimiento de sistemas de datos.
- **Machine Learning Engineer:** Desarrollador especializado en la implementación de modelos de aprendizaje automático.

### Tip

Para destacar en la Ingeniería de Datos, no te limites a aprender las habilidades técnicas; también trabaja en tus habilidades de comunicación y colaboración. La capacidad de traducir los requisitos de negocio en soluciones de datos efectivas es esencial.

**i** Actividad Práctica:

- **Explora Roles de Data Engineer:** Investiga ofertas de trabajo para Data Engineers en línea. Compara las responsabilidades y requisitos de distintas empresas para comprender la variabilidad en el rol.
- **Desarrolla tus Habilidades Técnicas:** Identifica una habilidad técnica relevante para Data Engineers que desees mejorar, como la programación en Python o el manejo de bases de datos SQL. Dedicar tiempo a practicar y fortalecer esa habilidad.

## 9 ¿Qué aprendimos?

En este tema, hemos explorado qué es un Data Engineer, destacando su rol central en la gestión y transformación de datos. Hemos identificado las responsabilidades clave, las habilidades necesarias y las oportunidades laborales en el campo de la Ingeniería de Datos. Además, enfatizamos la importancia de desarrollar habilidades técnicas y de comunicación para sobresalir en esta profesión en constante evolución.