# Data Engineer Roadmap for 2025

El creciente mar de datos exige profesionales capacitados para liberar su potencial. Impulsada por un asombroso **aumento del 21 % en las vacantes de trabajo** para profesionales de datos pronosticado por Zippia **entre 2018** y **2028** , la demanda de ingenieros de datos capacitados está en su nivel más alto. Los ingenieros de datos desempeñan un papel fundamental en la creación y el mantenimiento de la infraestructura que transforma los datos sin procesar en información procesable.

Si buscas una carrera gratificante en ingeniería de datos, esta hoja de ruta te proporciona las habilidades y tecnologías esenciales que necesitarás para navegar por el panorama dinámico de 2025. Esta hoja de ruta para ingenieros de datos describe un enfoque paso a paso, dividido en secciones sobre el desarrollo de habilidades de programación fundamentales, el dominio de las tecnologías de big data y las plataformas en la nube. ¿Quieres un camino claro hacia el éxito en ingeniería de datos? Únete al curso de ciencia de datos de Scaler para obtener una hoja de ruta completa y proyectos prácticos. Ya sea que seas un principiante absoluto o que busques mejorar tus habilidades, esta hoja de ruta proporciona una dirección clara para impulsar tu viaje en ingeniería de datos

MSc Ruben Quispe L.


# ¿Qué hace un ingeniero de datos?

Los ingenieros de datos son los arquitectos detrás de escena, los maestros constructores de la infraestructura de datos. Desempeñan un papel fundamental en el ecosistema de la ciencia de datos al diseñar, desarrollar y mantener los sistemas que recopilan, almacenan, transforman y hacen que los datos sean accesibles para su análisis.

MSc Ruben Quispe L.

A continuación, se presenta un análisis más detallado de las responsabilidades clave de un ingeniero de datos:

- Diseñar y automatizar canales de datos para mover datos desde diversas fuentes.
- Elegir y gestionar soluciones de almacenamiento de datos (bases de datos, almacenamiento en la nube).
- Garantice la precisión y la coherencia de los datos mediante controles de limpieza y calidad.
- Utilizar tecnologías de big data para el procesamiento de datos a gran escala.
- Colaborar con científicos y analistas de datos para traducir las necesidades de datos en soluciones.

# Requisitos previos para convertirse en ingeniero de datos

Si bien no existe un camino único y prescrito para convertirse en ingeniero de datos, una base sólida lo prepara para el éxito. Estos son los requisitos previos generales que debe tener en cuenta:

## Antecedentes educativos:

- Un **título de grado en informática, tecnología de la información** o un campo relacionado es un punto de partida común. Estos programas proporcionan una base sólida en lenguajes de programación como Python o Java, algoritmos como ordenación y búsqueda, y sistemas de bases de datos, incluidas bases de datos relacionales (SQL) y no relacionales (NoSQL). Sin embargo, algunos ingenieros de datos ingresan al campo con títulos en matemáticas, estadística o incluso física, aprovechando sus habilidades analíticas y de resolución de problemas.
- En el mundo actual, impulsado por los datos, puede resultar ventajoso contar con un máster en ciencia de datos o un programa especializado en ingeniería de datos. Estos programas ofrecen conocimientos profundos sobre tecnologías de big data, almacenamiento de datos y computación distribuida.
- 

### Pero recuerda

- No siempre es obligatorio tener un título universitario. Las personas autodidactas con sólidas habilidades demostrables y experiencia adquirida a través de cursos en línea, campamentos de entrenamiento o proyectos personales también pueden ingresar en este campo.

# Pasos para convertirse en ingeniero de datos: hoja de ruta

- Esta hoja de ruta proporciona a los aspirantes a ingenieros de datos las habilidades y tecnologías esenciales para prosperar en este campo dinámico. ¡Siga estos pasos y cultive el amor por el aprendizaje para convertirse en una estrella de la ingeniería de datos!

# Paso 1: Desarrollo de habilidades fundamentales (1 a 3 meses)

- Tu recorrido en ingeniería de datos comienza con una base sólida. Este paso se centra en las habilidades básicas de programación, los fundamentos de la informática y el lenguaje esencial para interactuar con bases de datos: SQL.

## Conceptos básicos de programación (Python):

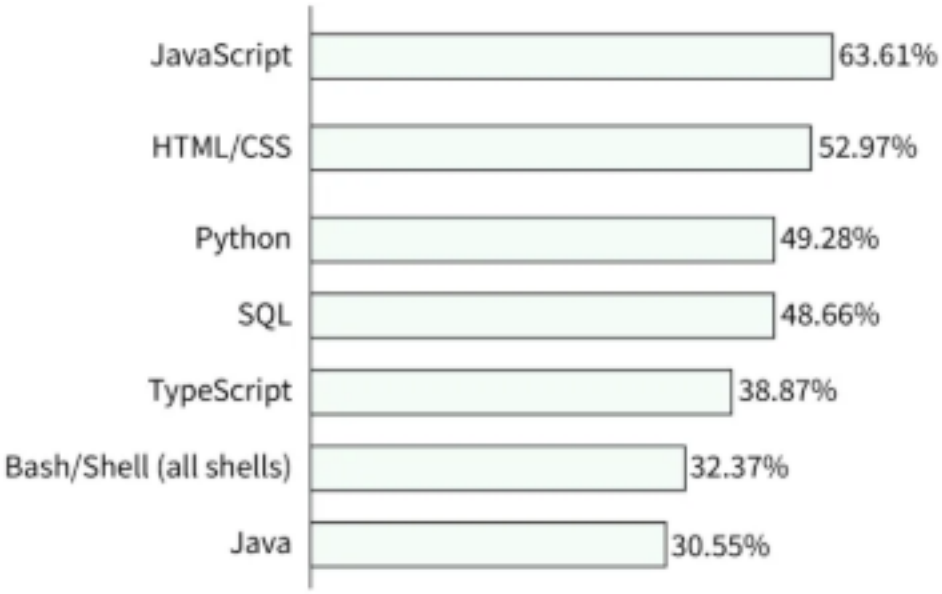
- Domine la sintaxis, las estructuras de datos, el flujo de control y las funciones de Python para manipular y analizar datos de manera eficaz. La versatilidad de Python lo convierte en una opción popular en ingeniería de datos.

## Fundamentos de informática:

- Comprenda conceptos básicos de informática como la gestión de memoria, los algoritmos y la complejidad de los datos. Esta base le ayudará a entender cómo se procesan y almacenan los datos en los sistemas informáticos.

## SQL (lenguaje de consulta estructurado):

- Aprenda a consultar bases de datos relacionales mediante SQL. Domine técnicas como **SELECT, JOIN** y **WHERE** para recuperar, filtrar y manipular datos almacenados en estas bases de datos.
- Según la Encuesta anual para desarrolladores de Stack Overflow de 2023 , SQL y Python se encuentran entre los lenguajes de programación más populares, con más del 49 % de los encuestados utilizándolos.




SCALER

### Recursos:

Este paso requiere una base sólida en varias áreas. A continuación, se ofrecen algunos recursos para comenzar:

- **Cursos en línea:** explore plataformas como Scaler para obtener diversos cursos sobre programación Python, fundamentos de informática y SQL.
- **Libros:** considere libros como "Automate the Boring Stuff with Python" de Al Sweigart, "Python Crash Course" de Eric Matthes, "Introduction to Algorithms" de Cormen et al., y "Cracking the Coding Interview" de Gayle Laakmann McDowell.
- **Tutoriales:** Utilice plataformas interactivas como W3Schools SQL Tutorial y SQLBolt para practicar la escritura de consultas SQL.

## Paso 2: Exploración de diferentes tipos de bases de datos (1-2 meses)

A medida que profundiza en la ingeniería de datos, comprender los distintos tipos de bases de datos se vuelve crucial. Este paso le brinda el conocimiento necesario para elegir la solución de almacenamiento adecuada para las diferentes necesidades de datos.

### Bases de datos relacionales (MySQL, PostgreSQL):

Estas bases de datos estructuradas almacenan datos en tablas organizadas con filas y columnas. Obtenga información sobre el diseño de esquemas, las técnicas de normalización y los métodos de consulta específicos de las bases de datos relacionales.

### Bases de datos NoSQL (MongoDB, Cassandra):

Las bases de datos NoSQL ofrecen flexibilidad para almacenar conjuntos de datos no estructurados o de gran tamaño. Explore los diferentes tipos de bases de datos NoSQL (documento, clave-valor, etc.) y comprenda sus ventajas y casos de uso.

### Almacenamiento de datos (Amazon Redshift, Google BigQuery):

Los almacenes de datos son repositorios especializados diseñados para almacenar datos históricos optimizados para el análisis. Conozca los conceptos de almacenamiento de datos (procesos ETL, modelado de datos) y explore soluciones de almacenamiento de datos basadas en la nube.

### Recursos:

En este paso se exploran diversas tecnologías de bases de datos. A continuación, se ofrecen algunos recursos para mejorar su aprendizaje:

- **Cursos en línea:** Plataformas como Scaler ofrecen cursos sobre bases de datos relacionales, bases de datos NoSQL y almacenamiento de datos.
- **Tutoriales:** utilice tutoriales de W3Schools o la documentación oficial de la base de datos (por ejemplo, MongoDB Getting Started) para adquirir experiencia práctica.

¡Pasemos al paso 3, donde te adentrarás en el mundo del procesamiento de datos!

## Paso 3: Dominar el procesamiento de datos (2-3 meses)

La transformación de datos es el núcleo de la ingeniería de datos. Este paso le proporciona los conocimientos y las herramientas para manipular y preparar los datos para su análisis.

### ETL (Extraer, Transformar, Cargar):

ETL es un proceso de ingeniería de datos fundamental para trasladar datos desde diversas fuentes a un almacén o lago de datos. Conozca las diferentes etapas de ETL (extracción, transformación, carga) y explore las herramientas que se utilizan para la integración de datos.

### Procesamiento por lotes y en streaming:

El procesamiento por lotes se ocupa de grandes conjuntos de datos en intervalos específicos, mientras que el procesamiento en tiempo real maneja flujos de datos continuos. Comprenda los conceptos y elija el enfoque adecuado en función del volumen de datos y las necesidades de procesamiento.

Mejore sus habilidades: muchas plataformas en línea ofrecen ejercicios y proyectos prácticos para consolidar su comprensión del procesamiento de datos.

### Recursos:

Este paso se centra en las técnicas de procesamiento de datos. A continuación, se ofrecen algunos recursos para reforzar sus conocimientos:

- **Cursos en línea:** explore plataformas en línea que ofrecen cursos sobre procesos ETL, procesamiento por lotes y procesamiento de transmisión.
- **Libros:** considere libros como “Ingeniería de datos: construcción de sistemas de análisis escalables” de Matt Leahu y “Procesamiento de datos en tiempo real con Apache Flink” de Fabian Hueske para un aprendizaje en profundidad.

## Paso 4: Exploración de tecnologías en la nube (1-2 meses)

El volumen de datos en constante crecimiento requiere soluciones escalables y rentables. Este paso le presenta las plataformas de computación en la nube que desempeñan un papel fundamental en la ingeniería de datos moderna.

### Introducción a las plataformas de computación en la nube (AWS, Google Cloud):

Las plataformas en la nube ofrecen acceso a demanda a recursos informáticos, almacenamiento y bases de datos. Explore los principales servicios en la nube, como computación, almacenamiento y bases de datos, que ofrecen los principales proveedores, como Amazon Web Services (AWS) y Google Cloud Platform (GCP). Obtenga una comprensión básica de cómo se pueden aprovechar estas plataformas para las tareas de ingeniería de datos.

Domine la nube: muchos proveedores de nube ofrecen niveles gratuitos o pruebas para que pueda experimentar y obtener experiencia práctica.

### Recursos:

Este paso presenta plataformas de computación en la nube para ingeniería de datos. A continuación, se incluyen algunos recursos para comenzar:

- **Documentación del proveedor de la nube:** explore la documentación oficial de AWS y Google Cloud para comprender sus servicios de nube en detalle.

- **Cursos en línea:** explore plataformas en línea que ofrecen cursos sobre los fundamentos de la computación en la nube y proveedores de nube específicos (AWS, GCP).

# Paso 5: Aprendizaje de tecnologías de Big Data (2-3 meses)

A medida que los volúmenes de datos se disparan, los métodos de procesamiento tradicionales tienen dificultades para seguir el ritmo. Este paso le presenta las tecnologías de big data que le permiten gestionar conjuntos de datos masivos de manera eficiente.

## Ecosistema Hadoop:

Hadoop es un marco de trabajo fundamental para procesar y almacenar grandes conjuntos de datos en clústeres de computadoras. Explore componentes centrales como HDFS (almacenamiento distribuido), YARN (gestión de recursos) y MapReduce (paradigma de procesamiento de datos). Comprender el ecosistema de Hadoop le permitirá adquirir conocimientos sobre un marco de trabajo de procesamiento de big data ampliamente utilizado.

## Apache Spark:

Spark es un popular marco de código abierto para el procesamiento de datos a gran escala, que suele utilizarse junto con Hadoop. Conozca las funcionalidades de Spark (Spark SQL, Spark Streaming) y sus ventajas sobre el MapReduce tradicional, como las capacidades de procesamiento en memoria para un rendimiento más rápido.

## Recursos:

Este paso se adentra en las tecnologías de big data. A continuación, se ofrecen algunos recursos valiosos que le ayudarán a avanzar:

- **Cursos en línea:** explore plataformas en línea para cursos sobre el ecosistema Hadoop y Apache Spark.
- **Libros:** considere libros como “Hadoop: The Definitive Guide” de Tom White (una guía completa) o “Learning Spark” de Holden Karau et al. (una introducción práctica).
- **Tutoriales:** utilice los tutoriales y la documentación del sitio web de Apache Spark para adquirir experiencia práctica.

# Paso 6: Desarrollar habilidades de gestión de datos (2 a 4 meses)

Una vez que domine las tecnologías fundamentales, estará listo para sumergirse en el corazón de la ingeniería de datos: la creación de canales de datos. Este paso le proporcionará las habilidades prácticas para diseñar, desarrollar y mantener flujos de trabajo automatizados que trasladan los datos desde el origen hasta el destino.

## Desarrollo de habilidades de canalización de datos:

Aprenda a diseñar canales de datos que extraigan datos de diversas fuentes (bases de datos, API, extracción de datos web), los transformen en un formato utilizable (limpieza, filtrado) y los carguen en almacenes o lagos de datos para su análisis. Explore herramientas populares de orquestación de canales de datos como Apache Airflow, Luigi y Prefect.

Práctica: la mejor manera de consolidar sus habilidades de canalización de datos es a través de la experiencia práctica. Considere trabajar en proyectos personales o contribuir a proyectos de canalización de datos de código abierto para desarrollar su cartera y mostrar sus capacidades.

## Recursos:

Este paso se centra en el desarrollo práctico de los flujos de datos. A continuación, se incluyen algunos recursos para potenciar su proceso de aprendizaje:
**Cursos en línea:** consulte las plataformas en línea que ofrecen cursos sobre desarrollo de canalizaciones de datos y herramientas de orquestación populares.
Obtenga una formación integral sobre flujos de datos con el curso de ciencia de datos de Scaler . Aprenda de los expertos de la industria y gane experiencia práctica.

- **Tutoriales:** explore los tutoriales y la documentación proporcionada por herramientas de orquestación de canalizaciones de datos como Apache Airflow, Luigi y Prefect.
- **Proyectos de código abierto:** busque proyectos de canalización de datos de código abierto aptos para principiantes en plataformas como GitHub para contribuir y adquirir experiencia práctica.

# Paso 7: Adquirir experiencia práctica y aplicarla (el tiempo varía)

Una vez que hayas establecido una base sólida en conceptos y tecnologías de ingeniería de datos, es hora de consolidar tu aprendizaje mediante la experiencia práctica. Este paso ofrece sugerencias de proyectos categorizadas por nivel de dificultad para ayudarte a perfeccionar tus habilidades y crear un portafolio atractivo.

## Proyectos para principiantes (1-2 meses):

- **Cómo crear un raspador web simple:** practique la extracción de datos escribiendo un script de Python para extraer datos de un sitio web (por ejemplo, información de productos, datos meteorológicos). Utilice bibliotecas como Beautiful Soup o Scrapy.
- **Desafío de limpieza y transformación de datos:** encontrar un conjunto de datos disponible públicamente (por ejemplo, datos gubernamentales, portales de datos abiertos) y practicar técnicas de limpieza de datos (manejo de valores faltantes, inconsistencias de formato) y métodos de transformación de datos (creación de nuevas características, normalización de datos) utilizando bibliotecas de Python como Pandas y NumPy.
- **Creación de una canalización de datos básica:** diseñe y desarrolle una canalización de datos sencilla utilizando una herramienta como Apache Airflow. Esto podría implicar extraer datos de un archivo CSV local, realizar transformaciones básicas y cargarlos en una base de datos como SQLite.

## Proyectos intermedios (2-4 meses):

- **Análisis de datos de sensores:** explore el análisis de datos en tiempo real mediante la simulación de datos de sensores (por ejemplo, lecturas de temperatura) y la creación de una secuencia de datos en tiempo real con Apache Spark Streaming. Visualice los datos en tiempo real con una herramienta de creación de paneles como Apache Kafka.
- **Creación de un motor de recomendaciones:** aproveche un conjunto de datos de películas u otro conjunto de datos de su elección para crear un motor de recomendaciones simple mediante técnicas de filtrado colaborativo. Explore bibliotecas como scikit-learn para implementar algoritmos de recomendaciones.

- **Desarrollo de un almacén de datos basado en la nube:** utilice una plataforma en la nube como AWS o Google Cloud para configurar un almacén de datos. Extraiga datos de varias fuentes, transfórmelos y cárguelos en el almacén de datos en la nube para su análisis.

### Proyectos avanzados (más de 4 meses):

- **Creación de un pipeline de aprendizaje automático:** combine sus habilidades de ingeniería de datos con el aprendizaje automático desarrollando un pipeline de datos completo para un proyecto de aprendizaje automático. Esto podría implicar preprocesamiento de datos, ingeniería de características, entrenamiento de modelos y evaluación mediante herramientas como TensorFlow o PyTorch.
- **Panel de análisis en tiempo real:** desarrolle un panel de análisis en tiempo real que visualice datos de una fuente de transmisión (por ejemplo, redes sociales, datos del mercado de valores). Utilice herramientas como Apache Kafka y Apache Flink para procesar los datos de transmisión y una biblioteca de visualización como Plotly o Dash para crear paneles interactivos.
- **Análisis de big data con Apache Spark:** trabaje con un gran conjunto de datos (por ejemplo, datos meteorológicos públicos, datos de redes sociales) y aproveche Apache Spark para el procesamiento distribuido de datos. Realice tareas complejas de análisis de datos, como agregaciones a gran escala, detección de anomalías o análisis de sentimientos.

# Habilidades avanzadas requeridas para un ingeniero de datos: mejore su experiencia

A medida que avanza en su recorrido de ingeniería de datos, considere perfeccionar estas habilidades avanzadas para elevar su experiencia y abordar desafíos aún más complejos:

- **Fundamentos del aprendizaje automático (ML)** : si bien los ingenieros de datos no necesariamente crean modelos de ML de producción, es valioso comprender los conceptos básicos de ML, como algoritmos, métricas de evaluación de modelos e ingeniería de características. Este conocimiento le permite colaborar de manera eficaz con científicos de datos y crear canales de datos sólidos para proyectos de aprendizaje automático.
- **Experiencia en la nube (AWS, GCP, Azure)** : las plataformas en la nube son la base de la infraestructura de datos moderna. Un conocimiento profundo de un proveedor de nube específico (AWS, Google Cloud Platform, Microsoft Azure) le permite aprovechar sus servicios administrados para el almacenamiento, procesamiento y análisis de datos. Esta experiencia agiliza las tareas de ingeniería de datos y garantiza la escalabilidad.
- **Seguridad y gobernanza de los datos** : la seguridad de los datos es primordial. Los ingenieros de datos deben comprender los controles de acceso a los datos, las técnicas de cifrado y las normas de cumplimiento para garantizar la privacidad y la seguridad de los datos dentro de los canales de datos que crean.
- **Sistemas distribuidos y DevOps** : la ingeniería de datos suele implicar trabajar con sistemas distribuidos que procesan datos en varias máquinas. Comprender los conceptos de sistemas distribuidos (tolerancia a fallas, escalabilidad) y los principios de DevOps

(integración continua/entrega continua) le permite crear e implementar canales de datos sólidos de manera eficiente.

- **Marcos de procesamiento de big data (Spark, Flink)** : más allá de Hadoop, dominar marcos como Apache Spark y Apache Flink lo equipa para manejar el procesamiento de datos en tiempo real y tareas complejas de análisis de datos en conjuntos de datos masivos de manera eficiente.
- **Sistemas de transmisión y mensajería de datos (Kafka, Kinesis)** : las canalizaciones de datos en tiempo real requieren herramientas especializadas. Obtenga información sobre plataformas de transmisión de datos como Apache Kafka o AWS Kinesis para procesar y gestionar transmisiones de datos de alta velocidad.
- **Herramientas de visualización de datos (Tableau, Power BI):** si bien los científicos de datos suelen tomar la iniciativa en la visualización de datos, los ingenieros de datos deben poseer habilidades básicas de visualización de datos utilizando herramientas como Tableau o Power BI. Esto le permite comunicar información de datos de manera eficaz a audiencias técnicas y no técnicas.