

Semana 1

Introducción al Ecosistema de Datos

Bootcamp de Data Engineering
Ian Saura

Bienvenidos

Ian Saura – Instructor:

- **Ingeniero Biomedico**
- **+4 años de experiencia como Data Engineer**
- **Profesor de infraestructura para ciencia de datos en UNSAM**

Sobre el Bootcamp: Curso intensivo enfocado en la práctica. Se recorrerá el concepto de pipeline de datos de extremo a extremo usando herramientas reales

Dinámica de clases: Clases en vivo con ejemplos prácticos, participación activa, espacio para preguntas y comunidad de soporte en Discord.



Objetivos de la clase

Conocer el ecosistema de datos: Familiarizarse con los componentes, roles y herramientas clave en ingeniería de datos.



Entender el flujo de trabajo: Aprender cómo se conectan las etapas desde la ingesta hasta la visualización de datos.



Introducción al proyecto final: Presentación del proyecto integrador que se desarrollará a lo largo del bootcamp.



¿Qué hace un Data Engineer?

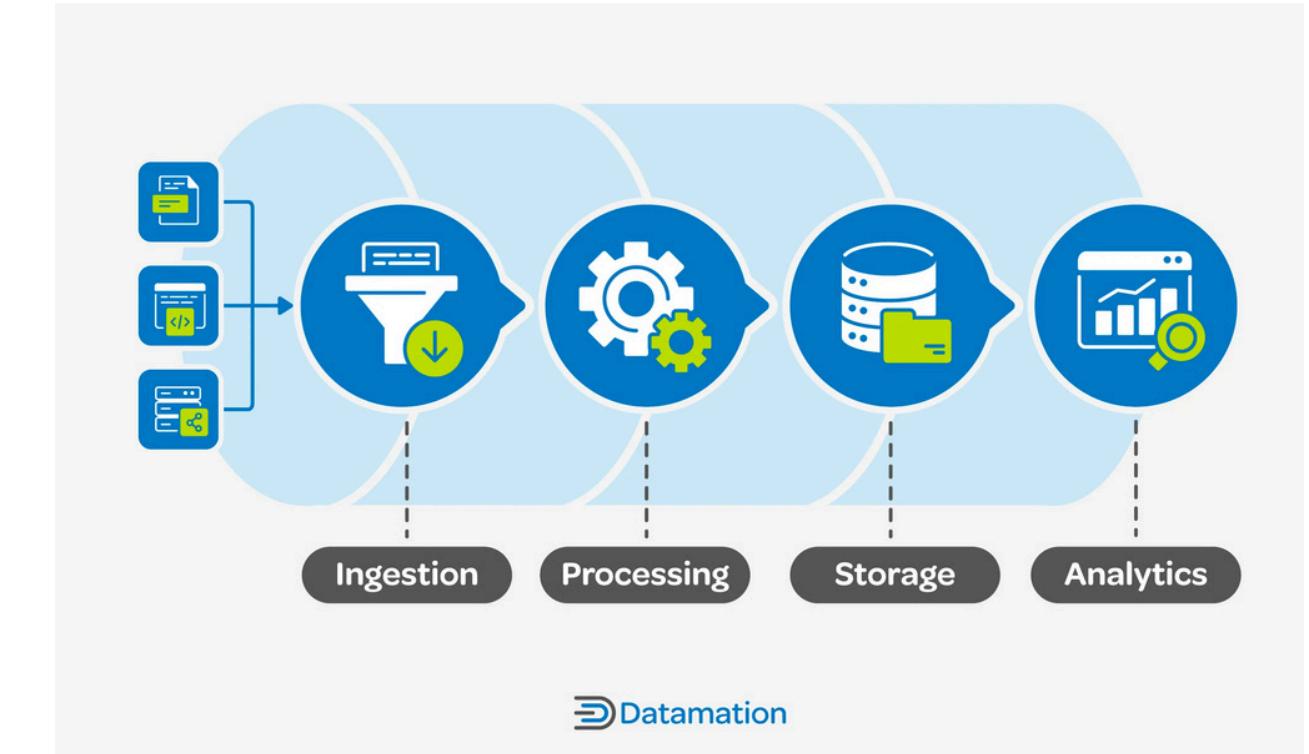


Responsabilidades centrales:
Ingesta, transformación,
modelado, orquestación y
aseguramiento de calidad de
los datos.

Diferencias con otros roles:
Data Engineer construye
infraestructura; Analyst genera
insights; Scientist desarrolla
modelos predictivos.

Visual comparativa: Mapa
simple de responsabilidades
entre Data Engineer, Analyst y
Scientist.

El flujo de trabajo de datos extremo a extremo



Fases clave del pipeline:

Ingesta → Almacenamiento → Transformación → Exposición. Cada etapa es crítica para la calidad final.

Visualización del flujo: Diagrama simple para ilustrar cómo viajan los datos desde su origen hasta un dashboard.

Ejemplo: datos de ventas: De una API de e-commerce, a un modelo limpio, y finalmente a un reporte interactivo.

Stop 1: Construyamos el flujo juntos

Allow Miro cookies

To offer you the best experience with embedded Miro boards we need your permission to access Miro cookies.

[Allow cookies](#)

[Not now](#)

Data as a Product

Pensar en datos como productos

Cada dataset debe tener dueño, ser versionado, documentado y mantenido.

Importancia de la calidad
Los datos deben ser confiables, consistentes y reproducibles para habilitar decisiones.

Ejemplo aplicado
Un dataset de métricas e-commerce versionado con documentación y validaciones automáticas.

Tipos de fuentes de datos

Bases de datos relacionales

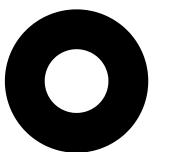
Fuentes estructuradas como PostgreSQL o MySQL, típicas en sistemas OLTP.

APIs y archivos planos
APIs REST o archivos CSV/JSON. Flexibles y comunes para integraciones.

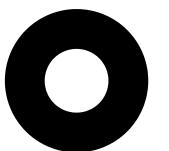
Streams de datos
Datos en tiempo real desde plataformas como Kafka o Kinesis.

Procesamiento de datos

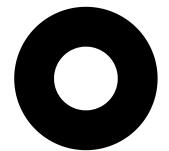
Procesamiento por lotes: Ejecución en intervalos definidos. Ideal para grandes volúmenes no urgentes. Ej: pandas, dbt.



Procesamiento en tiempo real: Análisis inmediato conforme llegan los datos. Ej: Spark Streaming, Kafka.



Aplicaciones comunes: Batch para reportes diarios; real time para alertas y dashboards actualizados.



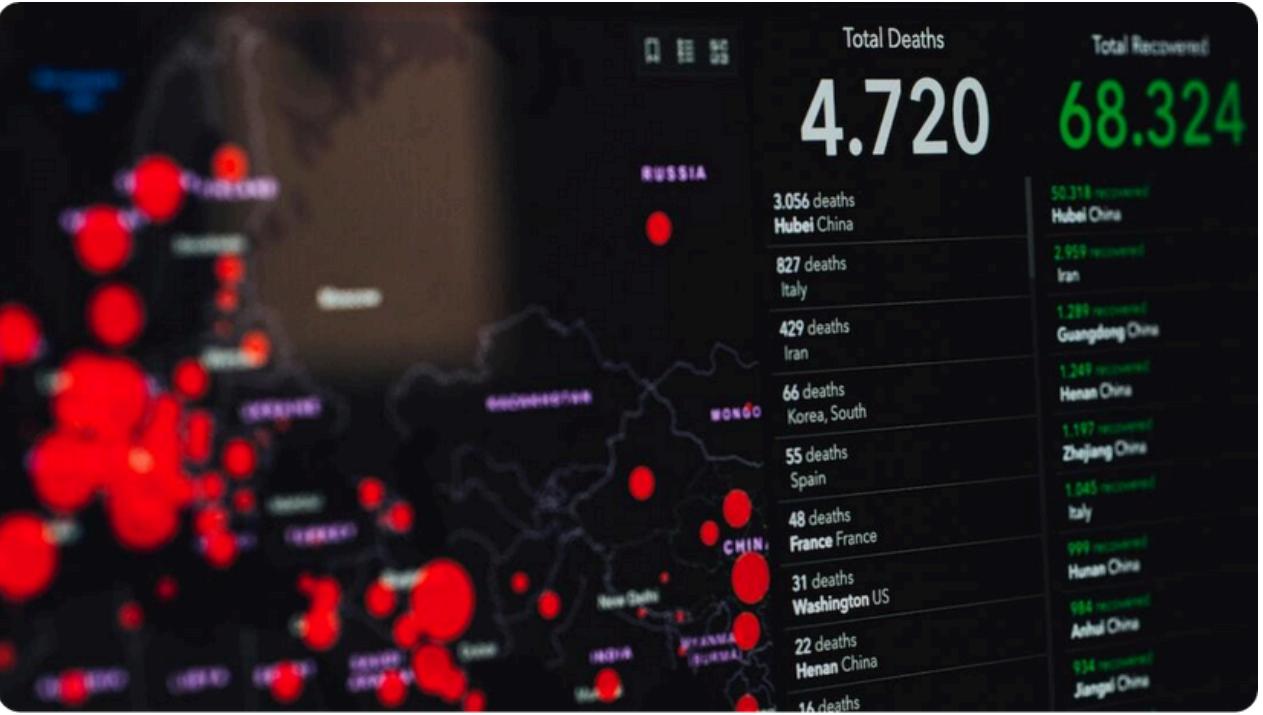


¿Dónde entra el Data Engineer en todo esto?

Rol transversal: Participa en todas las etapas: desde la ingesta hasta la exposición de datos.

Interacción con otros perfiles: Colabora estrechamente con analistas, científicos y equipos de producto.

Mapa de intervención: Visualización del pipeline con puntos de responsabilidad del Data Engineer.



Herramientas del ecosistema

- Ingesta: Airbyte, Python, APIs: permiten traer datos desde múltiples fuentes hacia nuestros sistemas.
- Transformación y almacenamiento: pandas, SQL, dbt para modelado; S3 y PostgreSQL como destino estructurado.
- Orquestación y BI: Airflow, Prefect automatizan flujos; Tableau, Metabase permiten visualización final.



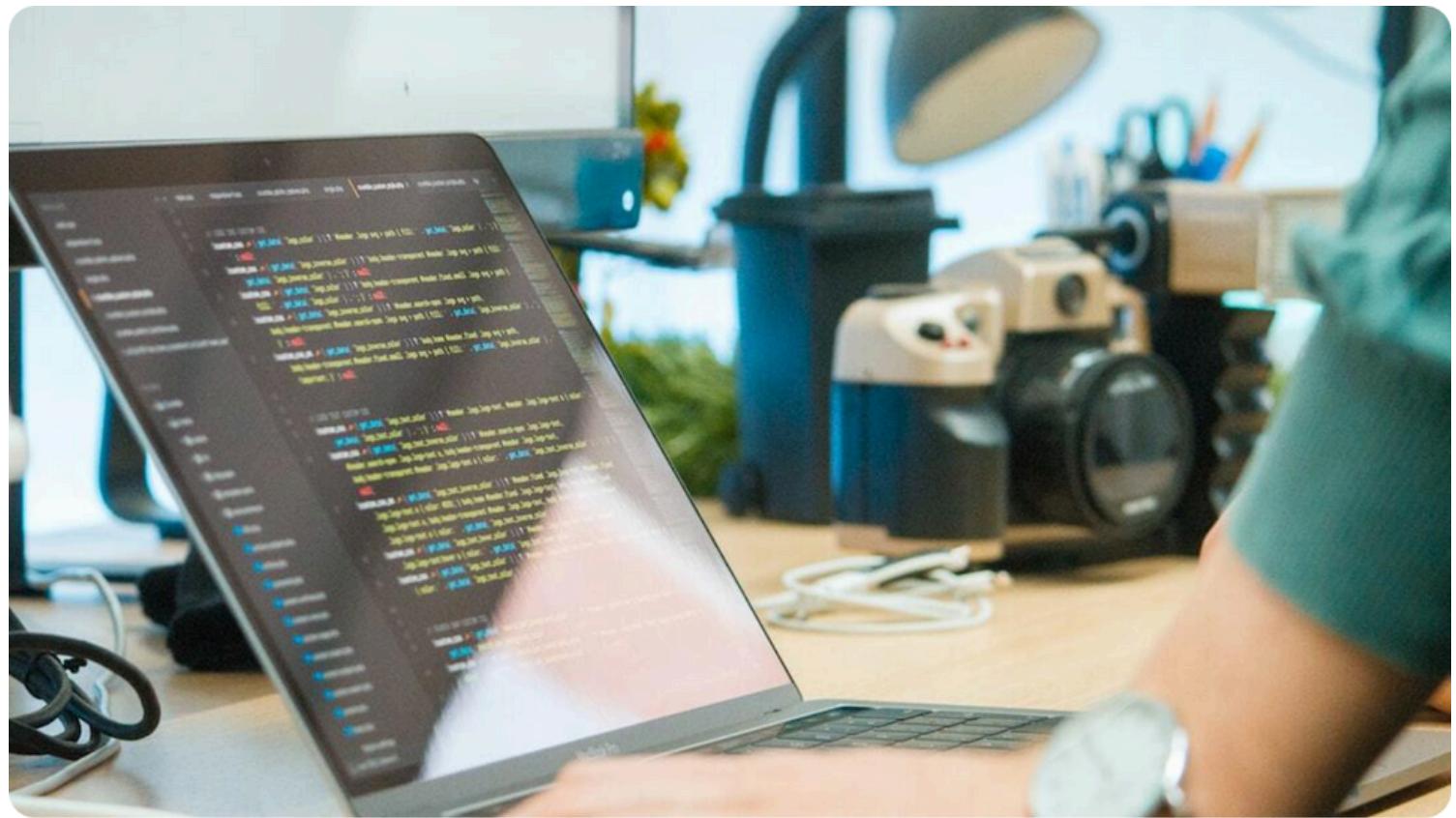
Pasemos a VS Code: nuestro primer ETL automatizado

- Objetivo: transformar un CSV y guardar la versión limpia.
- Herramientas: Python + pandas.
- Automatización: programar la ejecución con cron.
- Salida esperada: CSV limpio guardado con fecha.



¿Qué vamos a usar en el bootcamp?

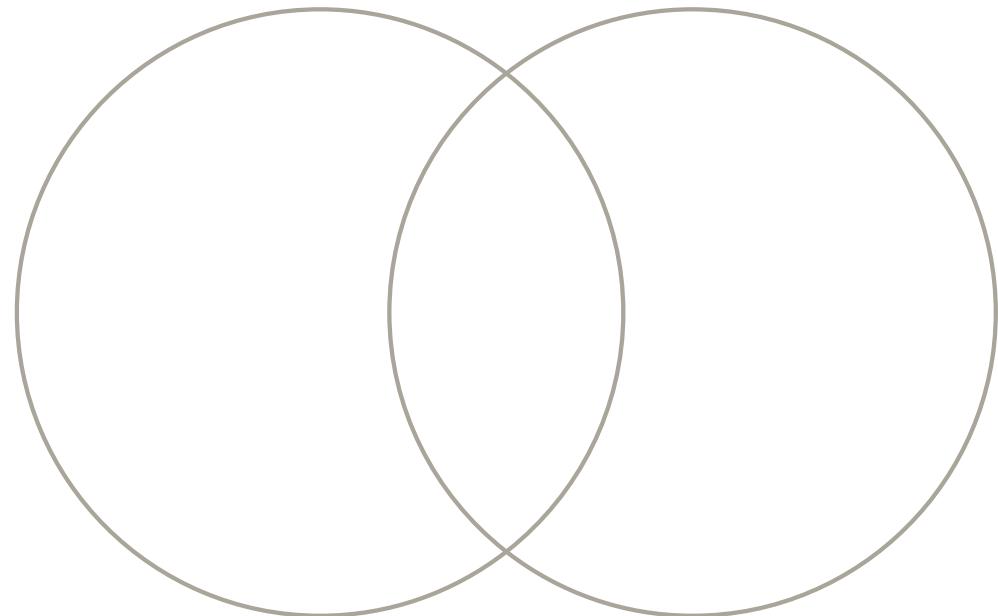
- Herramientas específicas: Python, pandas, SQL, DuckDB, dbt, cron/schedule (sin Docker ni servicios en la nube)
- Son livianas y fáciles de instalar, ideales para arrancar sin dolores de cabeza técnicos
- Permiten aprender los principios de pipelines sin depender de infraestructuras complejas
- Preparan para herramientas más avanzadas como Airflow sin necesidad de cambios bruscos



Habilidades clave que vas a desarrollar

- Pensamiento en pipelines: Aprender a construir flujos de datos eficientes, reutilizables y escalables.
- Modelado y automatización: Diseño de modelos lógicos de datos y automatización con orquestadores
- Buenas prácticas técnicas: Limpieza, testing, documentación en SQL y Python usando estándares profesionales.

Metodología de aprendizaje



Equilibrio teoría-práctica

Contenido técnico claro acompañado de ejercicios y casos reales para aplicar conceptos.

Proyectos semanales

Cada semana incluye entregas que construyen paso a paso el proyecto final.

Comunidad y feedback

Acompañamiento continuo por Discord y revisión de entregas con feedback del instructor.

¿Cómo será la evaluación?

- Proyecto final integrador: Construcción de un pipeline completo desde una fuente hasta visualización.
- Entregas semanales: Actividades prácticas que refuerzan los contenidos vistos en cada clase.
- Feedback continuo: Evaluación formativa de pares e instructor para mejorar iterativamente.

Proyecto final (vista general)

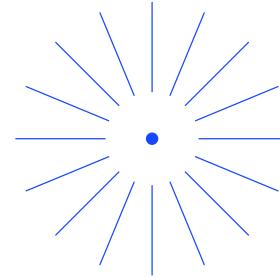
Pipeline completo

Desde CSVs reales hasta un dashboard: limpieza, modelado, transformación y visualización.

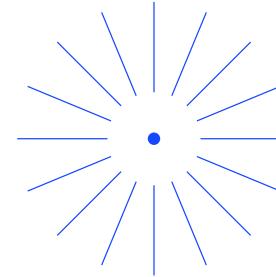
Caso de e-commerce
Dataset de ventas, productos y clientes. Métricas clave de negocio como ingreso y retención.

Herramientas integradas
Uso de Python, pandas, SQL, dbt, DuckDB, Cron para cubrir todas las etapas.

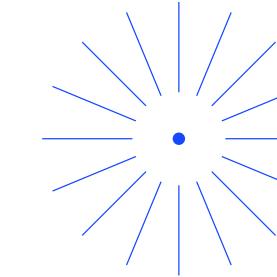
¿Qué necesito saber ahora?



SQL intermedio:
Capacidad de escribir
joins, subconsultas,
funciones de
agregación y filtros
complejos.



Python con pandas:
Manipulación de
DataFrames, limpieza,
transformaciones
básicas y
lectura/escritura de
archivos.

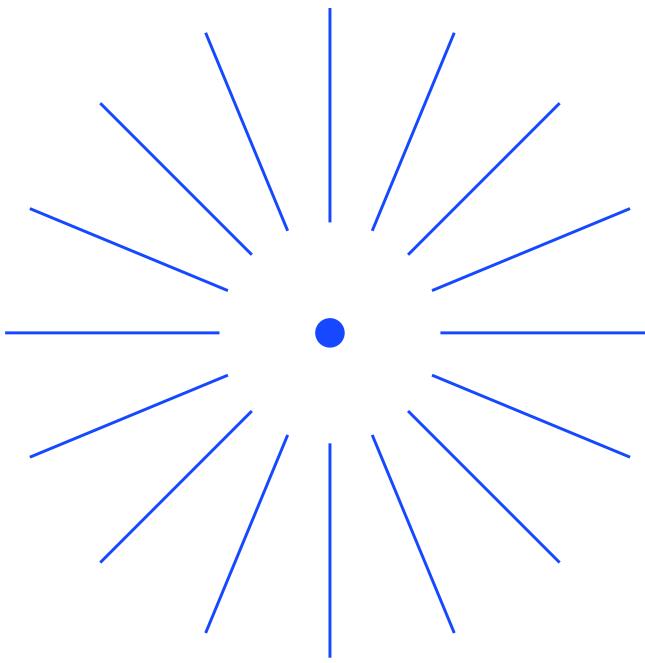


Actitud y comunidad:
Mentalidad analítica,
proactividad y
participación activa en
la comunidad de
Discord.

Recapitulación



- Conceptos clave: Ecosistema de datos, flujo extremo a extremo, rol del Data Engineer.
- Herramientas vistas: Airbyte, dbt, DuckDB, Cron, Tableau, entre otras.
- Primer entregable: Mapa de pipeline de datos de una empresa ficticia.



Preguntas y cierre

Espacio abierto
Tiempo para resolver dudas y
consultas en vivo con el instructor.



Ejercicios prácticos - Semana 1

Para entregar antes de la próxima clase

Ejercicio 1

Diagrama de flujo de datos: Diseñar un flujo de datos de una empresa cotidiana, identificando fuentes, almacenamiento, transformación y visualización.

Ejercicio 2

Glosario de herramientas: Investigar 6 herramientas vistas hoy, explicando su uso, ubicación en el pipeline y nivel de complejidad.

Ejercicio 3

Reto de roles: Comparar los perfiles de Data Engineer, Analyst, Scientist y Analytics Engineer. ¿Cuál te interesa más?

