1. Nombre de la materia:
Introducción a la Ingeniería de Datos.
2. Docente responsable:
a- Nombre y Apellido: Nicolás Tripp b- Máximo título alcanzado: Doctorado. c- Correo electrónico: tripio2000@gmail.com
3. Equipo docente:
-
4. Fechas:

Inicio: 01/09/2026	Finalización: 30/10/2026

5. Lugar de cursada:
Sede Rectorado - Modalidad: Híbrida. 16 clases virtuales + 2 clases presenciales de evaluación .
6. Días y horarios de cursada:
Martes y Jueves de 18 a 21 h.
7. Presentación de la materia:
En la ciencia contemporánea, la validez de una hipótesis no solo depende del modelo estadístico final, sino de la integridad del camino que recorrieron los datos desde su origen crudo hasta el gráfico final de la tesis. Este curso dota al estudiante de doctorado de las competencias necesarias para diseñar pipelines de datos reproducibles y robustos. Utilizando Python y Google Colab, el alumno aprenderá a automatizar la extracción, transformación, almacenamiento dimensional (OLAP) y reporte visual de sus datos de investigación, erradicando los procesos manuales propensos a errores.
8. Requisitos de admisibilidad:
Conocimientos básicos de programación. Conocimientos básicos de estadística descriptiva.
9. Duración en hs.
Horas teóricas: 36 Horas prácticas: 24 Horas totales: 60
10. Idioma del dictado:
Castellano.
11. ¿Podría dictarse una versión en idioma inglés?:
No.

12. Objetivos de aprendizaje:

Al finalizar el curso, el estudiante será capaz de:
1- Garantizar la trazabilidad: Diseñar un flujo de datos transparente donde cada gráfico e indicador sea auditable y replicable por terceros de principio a fin.
2- Ingestar y Estructurar: Extraer datos crudos de diversas fuentes y aplicar los principios de Tidy Data para su correcto análisis.
3- Modelar arquitecturas de datos (OLAP): Diseñar esquemas estrella y copo de nieve para estructurar datos complejos según sus dimensiones de análisis (tiempo, espacio, categorías).
4- Tratar anomalías con rigor científico: Diagnosticar y resolver valores nulos (NAs) y atípicos mediante imputación estadística metodológicamente justificada.
5- Comunicar con Impacto (Storytelling): Construir tableros interactivos y reportes visuales avanzados que traduzcan los datos en conocimiento científico directo para su audiencia.

13. Contenidos:

El Ecosistema de la Ciencia de Datos Reproducible. Revisión de fundamentos de la programación, estructura de datos y algoritmos. Entorno Google Colab y Python. Control de versiones; introducción a GIT.
Fundamentos de Ingesta y Estructura "Tidy". Sesión 3: Introducción a pandas.
Ingesta de datos crudos (CSV, Excel, JSON, APIs básicas). Definición de la unidad de análisis. Principios de Tidy Data: un caso por fila, una variable por columna. Transformación de matrices complejas a formatos planos (melt y pivot).
De datos sueltos a bases de datos. Integridad referencial. Definición de variables llave (Primary Keys y Foreign Keys). Cruzamiento y vinculación de datasets de orígenes distintos (merge y join). Fundamentos de almacenamiento de datos.
Diferencias analíticas y conceptuales entre sistemas transaccionales (OLTP) y sistemas analíticos (OLAP).
Diseño dimensional para Investigadores. Teoría del Modelado Dimensional.
Definición de Tablas de Hechos (Métricas/Variables dependientes) y Tablas de Dimensiones (Contextos/Variables independientes). Diseño práctico de un Esquema Estrella (Star Schema) y Esquema Copo de Nieve (Snowflake).
Diagnóstico y auditoría de datos. Análisis Exploratorio de Datos (EDA). Tipado estricto de variables en Python (numéricas, categóricas, strings, datetime).
Detección de fallas estructurales. Identificación de valores fuera del dominio teórico (edades negativas, fechas inconsistentes, textos corruptos). Análisis del impacto científico de los datos erróneos.
La ciencia de los datos faltantes. Tratamiento riguroso de registros vacíos (NULL / NA). Reglas de descarte: cuándo es metodológicamente aceptable eliminar una fila o columna completa y cuándo introduce sesgo de selección. Técnicas avanzadas de imputación. Imputación por tendencia central (media/moda), imputación estadística multivariable (KNN Imputer) e imputación cronológica para series de tiempo (backfill y forwardfill).
Anomalías, conglomerados y valores extremos. Identificación de Outliers.
Métodos estadísticos robustos (Rango Intercuartílico - IQR y Z-Score). Técnicas
de agrupamiento para etiquetado masivo y segmentación de datos atípicos antes
del reporte final.
Tableros de Control y Storytelling. Teoría de la Visualización Funcional.
Definición de la audiencia. Definición del mensaje central. Jerarquía visual y distribución de páginas/lienzos.
Visualización programática. Creación de gráficos de evolución (series de tiempo) y gráficos de distribución (histogramas, boxplots avanzados) usando Seaborn y Plotly interactivo dentro de Colab. Diseño de filtros interactivos (drill-down y roll-up) para permitir explorar las dimensiones del set de datos.

14. Trabajo de laboratorio:

El curso no contempla trabajo de laboratorio

15. Metodología de enseñanza:

Las clases virtuales serán teórico-prácticas. En particular, se realizarán varios trabajos prácticos guiados a lo largo de la cursada:
1. Trabajo Práctico 1: Principios de Tidy Data: un caso por fila, una variable por columna. Transformación de matrices complejas a formatos planos (melt y pivot).
2. Trabajo Práctico 2: Diseño práctico de un Esquema Estrella (Star Schema) y Esquema Copo de Nieve (Snowflake) basado en los datos reales de la investigación de los alumnos.
3. Trabajo Práctico 3: Técnicas avanzadas de imputación. Imputación por tendencia central (media/moda), imputación estadística multivariable (KNN Imputer) e imputación cronológica para series de tiempo (backfill y forwardfill).
4. Trabajo Práctico 4: Técnicas de agrupamiento para etiquetado masivo y segmentación de datos atípicos antes del reporte final.

16. Bibliografía obligatoria:

Taniar D., Rahayu W. (2021). Data Warehousing and Analytics. Fueling the Data Engine. Springer. DOI: 10.1007/978-3-030-81979-8
Vaisman, A., & Zimányi, E. (2014). Data warehouse systems: Design and implementation. Springer. https://doi.org/10.1007/978-3-642-54655-6
McKinney, W. (2022). Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.

17. Bibliografía complementaria:

-

18. Recursos didácticos para la enseñanza:

El curso tiene una duración de 60 horas: 48 horas virtuales en dos clases sincrónicas de 3 horas por semana; 6 horas presenciales para la evaluación, y 6 horas de trabajo individual. .
Las clases virtuales serán de carácter teórico-práctico. Las presentaciones teóricas serán acompañadas con ejemplos prácticos. Asimismo, se realizarán cuatro (4) trabajos prácticos guiados a lo largo de la cursada, tal como fuera explicado más arriba.
Un trabajo final integrará el contenido del curso y será la base de la evaluación.

19. Modalidad de evaluación:

Para aprobar el curso, el estudiante deberá defender un Proyecto Final individual dividido en dos componentes obligatorios:
1. El Entregable (50%): Entrega de un enlace a un cuaderno de Google Colab público, documentado mediante celdas de texto (Markdown) y completamente reproducible. El script debe tomar un dataset crudo de la disciplina del alumno, realizar el proceso completo de ETL, estructurarlo en un modelo dimensional (OLAP), resolver los NAs/atípicos de manera justificada, exportar la base de datos y crear un tablero básico con un mensaje claro.
2. La Defensa Académica (50%): Exposición oral e interactiva de 15 minutos. El alumno deberá demostrar en vivo el funcionamiento de su tablero (Looker Studio o Streamlit) enfocado en el Storytelling de su problema de investigación, justificando metodológicamente las decisiones tomadas durante la limpieza y el modelado de los datos.

20. Requisitos de aprobación:

Se calificará a los alumnos con una escala de 1 a 10, sobre la base de la evaluación descripta. El curso se aprueba con calificación 4.