

		UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO							
		FACULTAD DE INGENIERÍA							
TEMAS SELECTOS DE INGENIERÍA EN COMPUTACIÓN III				2596		10		8	
INTRODUCCIÓN A LA INGENIERÍA DE DATOS EMPRESARIAL									
Asignatura				Clave		Semestre		Creditos	
INGENIERÍA ELÉCTRICA				COMPUTACIÓN		INGENIERÍA EN COMPUTACIÓN			
División				Departamento		Carrera(s) en que se imparte			
Asignatura:				Horas/semana:				Horas/semestre:	
Obligatoria				Teóricas		4.0		Teóricas	
				Prácticas		0.0		Prácticas	
Optativa	X			Total		4.0		Total	
Modalidad:		Curso teórico							
Seriación obligatoria antecedente:				Ninguna					
Seriación obligatoria consecuente:				Ninguna					
Descripción del curso:									
Este curso proporciona a los estudiantes los conocimientos teóricos y prácticos para diseñar, construir y administrar pipelines de datos y arquitecturas escalables utilizando los servicios de Amazon Web Services (AWS). Los alumnos aprenderán a manejar el ciclo de vida completo de los datos, desde la ingesta y almacenamiento hasta la transformación, análisis y visualización, aplicando las mejores prácticas de la industria en un entorno de nube.									
Objetivo(s) del curso:									
Al finalizar el curso, los estudiantes serán capaces de:									
* Diseñar soluciones de datos robustas y escalables en la nube de AWS.									
* Implementar pipelines para la ingesta de datos en batch y en tiempo real (streaming).									
* Seleccionar y utilizar los servicios de almacenamiento de AWS más adecuados para diferentes casos de uso (Data Lakes, Data Warehouses).									
* Procesar y transformar grandes volúmenes de datos utilizando servicios serverless y de cómputo distribuido.									
* Orquestar, monitorear y optimizar flujos de trabajo de datos.									
* Aplicar principios de seguridad y gobierno de datos en la nube.									
Temario									
	NÚM.	NOMBRE				HORAS			
	1	Introducción al ámbito laboral en IT				4.0			
	2	Introducción al ámbito laboral en Data				4.0			
	3	Roles importantes en Data				4.0			

	4	Fundamentos de Cloud Computing	4.0			
	5	Fundamentos de Ingeniería de Datos	4.0			
	6	Almacenamiento y Gestión de Datos en AWS	4.0			
	7	Ingesta y Procesamiento de Datos en Batch	4.0			
	8	Ingesta y Procesamiento de Datos en Streaming	4.0			
	9	Data Warehousing & Análisis de Datos	4.0			
	10	Gobierno de Datos, Monitoreo & DataOps	12.0			
	11	Visualización de Datos	4.0			
	12	Proyecto Final - Desarrollo y Asesoría	12.0			
		Actividades prácticas	0.0			
		Total	64.0			

Prerequisitos: Se espera que el estudiante tenga conocimientos básicos en Python y SQL.

TEMARIO DESGLOSADO

1. Introducción al ámbito laboral en IT

En este módulo se establecen las bases que conforman un trabajo dentro del ámbito de tecnología.

Objetivo: Comprender como obtener un primer empleo, las responsabilidades que conlleva y las habilidades necesarias para desempeñarse correctamente en un entorno.

Contenido:

- 1- ¿Cómo obtener un primer empleo?
- 2- ¿Cómo realizar mi CV?
- 3- ¿Cómo generar un portafolio?

2. Introducción al ámbito laboral en Data

En este módulo se establecen las diferencias fundamentales entre un rol profesional en tecnología y un rol profesional en Data.

Objetivo: Comprender las diferencias entre distintos roles profesionales de la industria

Contenido:

- 1- ¿Cuáles roles existen en el ámbito laboral?
- 2- ¿Cuáles roles existen en Data?
- 3- ¿Cuáles son los diferentes tipos de industria en donde podemos trabajar en Data?

3. Roles importantes en Data

En este módulo profundizaremos sobre las diferencias en roles entre Científicos de Datos, Analistas de Datos, Ingenieros de Datos, entre otros roles relevantes en Data.

Objetivo: Comprender las diferencias entre los distintos roles profesionales en los perfiles de Data.

Contenido:

- 1- Profundización entre los roles de Data
- 2- ¿Cómo debo buscar trabajo en una empresa?
- 3- ¿Cómo identificar qué trabajos no concuerdan con el rol?

4. Fundamentos de Ingeniería de Datos

En estos módulos se establecen las bases conceptuales sobre la disciplina de ingeniería de datos, asegurando que todos los estudiantes compartan un entendimiento común antes de adentrarse en los servicios específicos de AWS.

Objetivo: Comprender el rol que cumple un ingeniero de datos en las organizaciones.

Contenido:

- 1- El rol del Ingeniero de Datos en el ciclo de vida de los datos.

5. Fundamentos de Cloud Computing

En este módulo se establecen las bases conceptuales del cómputo en la nube, asegurando que todos los estudiantes compartan un entendimiento común antes de adentrarse en los servicios específicos de AWS.

Objetivo: Generar un entorno de trabajo para un proyecto profesional.

Contenido:

- 1- Introducción a Cloud Computing (IaaS, PaaS, SaaS)
- 2- Conceptos clave de AWS: Regiones, Zonas de Disponibilidad, y servicios globales
- 3- Introducción a la Infraestructura como Código (IaC) con AWS CloudFormation o Terraform.
- 4- Fundamentos de seguridad en AWS: IAM (Roles, Usuarios, Políticas).
- 5- Laboratorio Práctico 1: Creación de una cuenta de AWS, configuración de un usuario IAM con permisos limitados y despliegue de una infraestructura básica con una plantilla de CloudFormation.

6. Almacenamiento y Gestión de Datos en AWS

Este módulo se centra en las soluciones de almacenamiento de AWS, pilar fundamental de cualquier arquitectura de datos. Los estudiantes aprenderán a elegir el servicio correcto según la naturaleza y el uso de los datos.

Objetivo: Comprender la importancia de un correcto almacenamiento de datos empresarial.

Contenido:

- 1- Amazon S3 (Simple Storage Service): El corazón del Data Lake.
- 2- Buckets, objetos, prefijos y políticas de ciclo de vida.
- 3- Clases de almacenamiento (Standard, Glacier, Intelligent-Tiering).
- 4- Seguridad en S3: Políticas de bucket y encriptación.

- 5- Data Lake vs. Data Warehouse: Conceptos y arquitecturas.
- 6- AWS Lake Formation: Creación y gestión segura de Data Lakes.
- 7- Bases de datos relacionales con Amazon RDS y Aurora.
- 8- Bases de datos NoSQL con Amazon DynamoDB.
- 9- Laboratorio Práctico 2: Construir un Data Lake básico en Amazon S3, ingestar datos de muestra y aplicar políticas de ciclo de vida y seguridad.

7. Ingesta y Procesamiento de Datos en Batch

Aquí los estudiantes se enfocarán en los patrones clásicos de ETL (Extract, Transform, Load) para procesar grandes volúmenes de datos de manera programada.

Objetivo: Analizar las necesidades de negocio para ofrecer una solución adecuada a sus necesidades.

Contenido:

- 1- Glue: Servicio serverless de ETL y catálogo de datos.
- 2- Crawlers para descubrir esquemas.
- 3- Desarrollo de trabajos ETL con PySpark en Glue Studio y notebooks.
- 4- Catálogo de datos y su integración con otros servicios.
- 5- Amazon EMR (Elastic MapReduce): Procesamiento de Big Data con clusters de Spark y Hadoop.
- 6- Orquestación de flujos de trabajo con AWS Step Functions.
- 7- Laboratorio Práctico 3: Crear un trabajo de ETL con AWS Glue que lea datos crudos desde S3, los transforme (limpieza, agregación) y los escriba en un formato optimizado (Parquet) en otra ubicación del Data Lake. Orquestar la ejecución con Step Functions.

8. Ingesta y Procesamiento de Datos en Tiempo Real (Streaming)

Este módulo introduce a los estudiantes en el mundo del procesamiento de datos en streaming, fundamental para casos de uso como el análisis de redes sociales, IoT o detección de fraude.

Objetivo: Ver una introducción al procesamiento de datos en tiempo real, comprender las diferencias fundamentales entre el procesamiento en Batch y entender su caso de uso.

Contenido:

- 1- Fundamentos del procesamiento de streaming (ventanas de tiempo, marcas de agua).
- 2- Amazon Kinesis:
 - a. Kinesis Data Streams: Para la ingesta de flujos de datos a gran escala.
 - b. Kinesis Data Firehose: Para entregar datos de streaming a destinos como S3 o Redshift de manera sencilla.
- 3- Kinesis Data Analytics: Para analizar datos en streaming con SQL o Flink.
- 4- AWS Lambda: Procesamiento de eventos en tiempo real y arquitecturas serverless.
- 5- Tarea 1: Desarrollar una aplicación de streaming que ingeste datos de un simulador (ej. datos de clics de una web) con Kinesis Data Streams, los procese con una función Lambda para enriquecerlos y los archive en S3 usando Kinesis Data Firehose.

9. Data Warehousing y Análisis de Datos

Una vez que los datos están limpios y almacenados, este módulo cubre cómo analizarlos para extraer valor y conocimiento.

Objetivo: Comprender la importancia del almacenamiento de los datos y las arquitecturas involucradas en esto a nivel empresarial.

Contenido:

- 1- Amazon Redshift: Data Warehouse en la nube a escala de petabytes.
- 2- Arquitectura (nodos líderes, nodos de cómputo).
- 3- Mejores prácticas de diseño de tablas (claves de distribución y ordenamiento).
- 4- Redshift Spectrum: para consultar datos directamente en el Data Lake de S3.
- 5- Amazon Athena: Consultas SQL interactivas y serverless sobre datos en S3.
- 6- Laboratorio Práctico 5: Cargar los datos transformados de los anteriores módulos en Redshift Serverless. Realizar consultas analíticas complejas. Usar Athena para consultar los datos directamente en S3 y comparar el rendimiento.

10. Gobierno, Monitoreo y Automatización (DataOps)

Estos módulos se enfocan en la operacionalización de las soluciones de datos, asegurando que sean seguras, confiables y mantenibles.

Objetivo: Comprender los conceptos de Gobierno, Monitoreo y Automatización y su importancia en las empresas.

Contenido:

- 1- Gobierno de Datos: Calidad, linaje y gestión de metadatos.
- 2- Monitoreo y Logging: Uso de Amazon CloudWatch (Logs, Métricas, Alarmas) y AWS CloudTrail.
- 3- CI/CD para Pipelines de Datos: Automatización de despliegues con AWS CodePipeline y AWS CodeBuild.
- 4- Optimización de costos en arquitecturas de datos.
- 5- Laboratorio Práctico 6: Configurar alarmas en CloudWatch para un pipeline de Glue. Crear un pipeline básico de CI/CD para automatizar el despliegue de una función Lambda.

11. Visualización de Datos

Este módulo presenta de manera introductoria la visualización de datos empresarial y su ayuda en la toma de decisiones.

Objetivo: Comprender la importancia de una correcta visualización de datos y los sesgos y errores críticos que esta puede generar en las tomas de decisiones.

Contenido:

- 1- Amazon QuickSight: Visualización de datos y Business Intelligence (BI).
- 2- Tableau Public
- 3- Laboratorio Práctico 7: Desarrollo de un tablero en Tableau Public.

12. Proyecto Final

Diseñar y construir un pipeline de datos de extremo a extremo en AWS que resuelva un problema de negocio propuesto.

Objetivo: Realizar un caso de negocio real que permita agregarlo como un proyecto de portafolio, aplicando todos los conocimientos adquiridos en el curso.

Contenido:

- 1- ¿Qué es una Solicitud de Propuesta, en qué consiste y para qué sirve?
- 2- ¿Qué es una Declaración de Trabajo, en qué consiste y para qué sirve?
- 3- ¿Qué es una Prueba de Concepto, en qué consiste y para qué sirve?
- 4- ¿Qué es un Producto Mínimo Viable, en qué consiste y para qué sirve?
- 5- ¿Cómo desarrollo un proyecto en equipo?
- 6- Asesoría para el desarrollo de proyectos finales.
- 7- Revisión y retroalimentación de avances.

Bibliografía básica	Temas para los que se recomienda
Reis, J., & Housley, M. (2022). <i>Fundamentals of data engineering: Plan and build robust data systems</i> . O'Reilly Media.	
Wittig, A., & Wittig, M. (2019). <i>Amazon Web Services in action (2nd ed.)</i> . Manning Publications.	
Chambers, B., & Zaharia, M. (2018). <i>Spark: The definitive guide - Big data processing made simple</i> . O'Reilly Media.	
Nussbaumer Knafllic, C. (2017). <i>Storytelling con datos: Visualización de datos para profesionales</i> . Anaya Multimedia. ISBN: 978-84-415-3930-3.	
Grus, J. (2019). <i>Data science from scratch: First principles with Python (2nd ed.)</i> . O'Reilly Media. ISBN: 978-1-492-04113-9	

Bibliografía complementaria	Temas para los que se recomienda
Kleppmann, M. (2017). <i>Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems</i> . O'Reilly Media.	
Matthes, E. (2019). <i>Python Crash Course: A Hands-On, Project-Based Introduction to Programming (2nd ed.)</i> . No Starch Press.	

Materiales del curso	
Cuenta activa de Google (Gmail, Google Colab)	
Acceso a internet	

--

Evaluación	
Proyecto	70%
Exámenes	10%
Prácticas	20%

Información del profesor

Nombre completo:	Caleb Aguilar Camargo
Correo electrónico institucional:	
Horario de la clase:	Lunes y Martes – 17:00 A 19:00 HRS
Semblanza corta del profesor.	Caleb Aguilar es un Ingeniero de Datos con más de 5 años de experiencia, cuya base profesional se asienta en una sólida formación como Lic. en Matemáticas. Su carrera se ha desarrollado principalmente en el ámbito de la consultoría, donde se ha enfocado en la entrega de soluciones de datos end-to-end y en la modernización de infraestructuras. Ha entregado valor a diversos sectores como Retail, Banca, Transporte y Fintech, trabajando para clientes globales de la talla de Kellogg's, CitiBanamex y Aeroméxico. Además de su rol técnico, tiene un gran interés en la pedagogía y la mentoría para las nuevas generaciones, una pasión que ha cultivado al dar clases en la UNAM y diversas charlas en eventos nacionales de divulgación y comunicación científica.