



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



Unidad de Competencia							
Ingeniería de Datos a Gran Escala							
Cuatrimestre	Fecha de elaboración			Modalidad	Tipo	Área de formación	
8	DD	MM	AÑO	• Escolarizada Opción: ○ Presencial	• Curso / Laboratorio	Programación e ingeniería de software	
	27	11	2025				
Perfil docente							
• Formación: ○ Ingeniería en Datos, Sistemas Computacionales, Software, Cloud o áreas equivalentes. ○ Posgrado deseable en Big Data, Sistemas Distribuidos, Ingeniería de Datos o Ciencia de Datos. • Experiencia profesional: ○ Mínimo 3 años trabajando con Spark, Databricks, Hadoop o ecosistemas distribuidos. ○ Conocimientos sólidos en ETL, pipelines, data lakes, data warehouses y procesamiento distribuido. • Dominio de inglés: ○ B2 o superior (documentación técnica y recursos cloud). • Competencias digitales: ○ Manejo de Spark (PySpark), Databricks Notebooks, MLflow, Delta Lake, Git, Cloud (AWS/Azure).							
Horas con Mediación Docente Cuatrimestrales		Horas con Mediación Docente semana/mes		Horas de Actividad Independiente	Horas de Práctica de Campo	Créditos	Academia
HT	HP	HT	HP	64	0	8	
32	32	2	2				
Horas de docencia total de la unidad de competencia: 64							
Propósito general:				Presentación:			
El estudiantado diseña y ejecuta pipelines de datos a gran escala				En la era del Big Data, el dominio de arquitecturas distribuidas es			



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



mediante Apache Spark y Databricks, aplicando arquitecturas distribuidas, optimización de consultas, control de versiones, integración con data lakes y despliegue de flujos de procesamiento orientados a análisis avanzado, ML y aplicaciones de IA.	esencial para cualquier ingeniero/a de datos o científico/a de datos. Esta unidad enseña los principios fundamentales del procesamiento masivo mediante Spark y su implementación moderna en Databricks, integrando conceptos esenciales como RDDs, DataFrames, Delta Lake, structured streaming, clusters y optimización. Se articula con MLOps, Cloud, Machine Learning II, Ingeniería de Características y Proyectos Integradores
Competencias	
Genéricas	Disciplinares
• Pensamiento sistémico. • Resolución de problemas complejos. • Trabajo colaborativo. • Comunicación técnica.	• Implementa procesamiento distribuido. • Diseña pipelines escalables con Spark. • Administra y optimiza clústeres de datos.
Profesionales	
• Domina Databricks para ingeniería de datos. • Integra Delta Lake para almacenamiento transaccional. • Aplica buenas prácticas en entornos productivos.	
Producto final	
Pipeline de Datos a Gran Escala, que incluya: Ingesta y transformación masiva de datos, Uso de Spark DataFrames, Optimización con Spark SQL y Delta Lake, Integración con MLflow / Databricks, Despliegue en clúster y documentación técnica.	
Enseñanza con perspectiva de género, interculturalidad y Derechos Humanos	
Con la intención de propiciar una enseñanza con perspectiva de género, interculturalidad y derechos humanos, la o el docente que imparta esta unidad de competencia deberá: • Eliminar y rechazar actitudes y comportamientos sexistas, clasistas, racistas y todo tipo de estereotipos que supongan discriminación y violencia en los ámbitos de enseñanza-aprendizaje (virtual y presencial), a partir del ejemplo en su práctica docente. • Promover que el desarrollo del programa educativo y material a utilizar se base en el respeto, protección, promoción y garantía de los derechos humanos, de la dignidad e integridad de todas las personas. • Promover la eliminación de estereotipos o cualquier otra práctica de discriminación o denigración de la persona, que pueda presentarse	



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



mediante el uso del lenguaje, el empleo de ejemplos, la diferenciación entre grupos, origen étnico o cualquier otra circunstancia.

- Respetar la identidad de género de las personas en todo momento y nombrarlas tal y como ellas lo deseen.
- Respetar y promover la igualdad de trato y de oportunidades entre las y los estudiantes, favoreciendo acciones afirmativas en beneficio de los grupos en situación de vulnerabilidad, incluso en las afectaciones personales provocadas por la pandemia COVID-19
- Promover y usar lenguaje incluyente y no sexista tanto en el salón de clases como en los trabajos, tesis y otros documentos que elabore la comunidad estudiantil.

Estos elementos son motivados por la información asequible en los programas de estudio de la Licenciatura en Derecho, Modalidad Educativa Presencial (sistema escolarizado), Tomo II, de la Universidad Nacional Autónoma de México (UNAM 2020).



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



Nivel Unidad/Subcompetencias	
Nombre de la Subcompetencia 1	Elementos de la subcompetencia 1
Introducción a Arquitecturas Distribuidas y Spark Core	Conocimientos: • Big Data: conceptos y desafíos • Sistemas distribuidos • Apache Spark: arquitectura, RDDs, ejecución distribuida • Clústeres, ejecutores y jobs Habilidades: • Uso de RDDs • Ejecución básica de jobs • Comprensión del DAG de ejecución Valores y actitudes: • Rigor técnico • Responsabilidad • Ética en manejo de datos a gran escala
Número de semanas	
4 semanas	
Propósito de la subcompetencia	
El estudiantado comprende las bases del procesamiento distribuido y ejecuta sus primeras tareas con Spark Core.	
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Laboratorio PySpark • Visualización de DAGs • Mini-casos prácticos	• Databricks Community • Spark UI • PySpark

Nombre de la Subcompetencia 2	Elementos de la subcompetencia 2
DataFrames, Spark SQL y Optimización Avanzada	Conocimientos: • DataFrames y optimizaciones
Número de semanas	



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



4 semanas	• Spark SQL • Catalyst Optimizer • Técnicas de particionado, caching y broadcast • Funciones de ventana Habilidades: • Consultas optimizadas • Transformaciones complejas • Estrategias de mejora de rendimiento Valores y actitudes: • Precisión • Mejora continua
Propósito de la subcompetencia	
El estudiantado transforma y consulta datos masivos utilizando DataFrames y Spark SQL de forma eficiente.	
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Laboratorio SQL • Optimización de queries • Análisis de rendimiento	• Spark SQL • Databricks Notebooks • Datasets de gran volumen

Nombre de la Subcompetencia 3	Elementos de la subcompetencia 3
Databricks, Delta Lake y Pipelines de Producción	Conocimientos: • Arquitectura Databricks • Delta Lake: ACID transactions, time travel • ETL/ELT escalable • MLflow + Spark • Integración con cloud
Número de semanas	
4 semanas	
Propósito de la subcompetencia	
El estudiantado implementa pipelines robustos y auditables utilizando Databricks y Delta Lake.	Habilidades:



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



	• Diseño de ETL distribuido • Gestión de versiones de tablas • Integración de MLflow en pipelines
	Valores y actitudes: • Profesionalismo • Responsabilidad operativa
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Laboratorio Delta Lake • Pipelines de transformación • Registro de modelos/pipelines	• Databricks • Delta Lake • MLflow

Nombre de la Subcompetencia 4	Elementos de la subcompetencia 4
Structured Streaming y Procesamiento en Tiempo Real	Conocimientos: • Structured Streaming • Fuentes y sinks (Kafka, sockets, Delta tables) • Procesamiento en micro-batches • Manejo de estado • Casos de uso en IA Habilidades: • Programación de flujos • Integración con Kafka/Delta • Monitoreo de pipelines Valores y actitudes: • Responsabilidad
Número de semanas	
4 semanas	
Propósito de la subcompetencia	
El estudiantado construye aplicaciones de streaming para procesar datos en tiempo real.	



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



	• Ética en tiempo real (datos sensibles)
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Laboratorio streaming • Creación de dashboards • Presentación del flujo	• Spark Streaming • Kafka (introductorio) • Databricks Jobs

Evaluación	
La evaluación de los aprendizajes se desarrollará de forma continua durante el proceso de enseñanza-aprendizaje a través de los siguientes momentos: Evaluación diagnóstica: Evaluación de ML II, Python y SQL. Evaluación formativa: Laboratorios progresivos, Optimización incremental, Miniproyectos, Retroalimentación continua Evaluación sumativa: Pipeline final de datos, Examen técnico, Presentación ejecutiva, Portafolio digital	
Estrategia de evaluación de los aprendizajes:	Instrumentos de Evaluación:
Estrategias que el estudiante desarrolla para la elaboración de las evidencias de aprendizaje. • Scripts PySpark • Pipelines Databricks • Informes técnicos • Métricas de rendimiento	• Rúbricas • Cotejos • Análisis de logs • Revisión de notebooks
Referencias	
<i>Las referencias consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
Básicas:	Complementarias:
• Databricks Academy Docs.	• O'Reilly – Learning Spark (2nd Edition).



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



- Apache Spark Documentation.
- Delta Lake Documentation

- MLflow Docs (Databricks).
- Papers de Spark y Delta Lake.

Videos:

Bases de Datos:

- CONRICYT
- Databricks Datasets
- Kaggle (Big Data collections)