

	Nombre del documento: Formato de Programa de Estudio de asignatura de Especialidad	Código: TecNM-AC-PO-007-02
		Revisión: 0
	Referencia a la Norma ISO 9001:2015 8.3, 8.3.1	Página 1 de 10

1. Datos Generales de la asignatura

Nombre de la asignatura:	Fuentes y técnicas de extracción de datos
Clave de la asignatura:	CDF-2502
SATCA¹:	3-2-5
Carrera:	Ingeniería en Sistemas Computacionales

2. Presentación

Caracterización de la asignatura Esta asignatura contribuye al perfil de egreso formando a los estudiantes en técnicas avanzadas de manejo, extracción y análisis de grandes volúmenes de datos. Los egresados serán capaces de aplicar estos conocimientos para resolver problemas complejos en diversas áreas, mejorando la toma de decisiones basadas en datos. En un mundo donde el análisis de datos es clave para la competitividad empresarial y la innovación científica, esta asignatura prepara a los estudiantes para trabajar en proyectos que requieran la extracción de información útil a partir de grandes volúmenes de datos, una habilidad fundamental en áreas como la inteligencia artificial, la ingeniería de software y el desarrollo de sistemas informáticos. La materia se enfoca en enseñar técnicas de extracción, transformación y carga (ETL) de datos, así como la utilización de herramientas de almacenamiento y manipulación de bases de datos no relacionales (NoSQL). Además, aborda el uso de lenguajes de programación como Python para la manipulación y análisis de datos. Esta asignatura se relaciona estrechamente con otras materias del módulo de especialidad en Ciencia de Datos, tales como Minería de Datos, Machine Learning y Big Data, ya que todas comparten el objetivo de dotar a los estudiantes de competencias específicas en la manipulación y análisis de datos. También se conecta con asignaturas como Bases de Datos y Algoritmos al requerir conocimientos previos en la estructuración de información y optimización de procesos computacionales.
--

¹ Sistema de Asignación y Transferencia de Créditos Académicos

	Nombre del documento: Formato de		Código: TecNM-AC-PO-007-02
	Programa de Estudio de asignatura de Especialidad		Revisión: 0
	Referencia a la Norma ISO 9001:2015	8.3, 8.3.1	Página 2 de 10

3. Intención didáctica.

Los contenidos deben abordarse de forma práctica y aplicada, vinculando constantemente los conceptos teóricos con ejemplos del mundo real. Los estudiantes deben aprender a aplicar las herramientas de extracción de datos en proyectos que simulen escenarios de la industria, permitiendo una comprensión más profunda y práctica del material.

El enfoque debe ser interdisciplinario, integrando el uso de tecnologías modernas como bases de datos NoSQL, lenguajes de programación como Python y herramientas de análisis de datos. Se priorizará la solución de problemas complejos mediante el uso de técnicas de extracción y procesamiento de datos, promoviendo el pensamiento crítico y la toma de decisiones basada en datos.

La extensión y profundidad deben ser equilibradas, profundizando en temas clave como la extracción, transformación y carga de datos (ETL), así como en las herramientas utilizadas para estas tareas. La asignatura debe garantizar que los estudiantes logren un dominio técnico que les permita aplicar los conceptos a diferentes tipos de fuentes de datos, tanto estructuradas como no estructuradas.

Se debe promover la participación activa en proyectos y trabajos colaborativos, con el objetivo de desarrollar competencias como el trabajo en equipo, la responsabilidad individual y la capacidad de investigación autónoma. También se incentivará el uso de plataformas colaborativas y la presentación de resultados en foros académicos.

Entre las competencias genéricas que se desarrollarán están:

- Resolución de problemas: Aplicación de métodos para enfrentar desafíos de extracción y análisis de datos.
- Trabajo en equipo: Colaboración efectiva en proyectos grupales.
- Aprendizaje autónomo: Capacidad de investigar nuevas herramientas y técnicas de extracción de datos por cuenta propia.
- Comunicación efectiva: Presentación clara y concisa de resultados y conclusiones a través de informes y exposiciones.

El docente debe asumir un rol de facilitador y guía, proporcionando las herramientas necesarias para que los estudiantes desarrollen sus propias soluciones. Además, debe fomentar la investigación y el aprendizaje autónomo, asegurando que los estudiantes sean capaces de adaptar y aplicar sus conocimientos a contextos cambiantes y nuevos problemas.

	Nombre del documento: Formato de	Código: TecNM-AC-PO-007-02
	Programa de Estudio de asignatura de Especialidad	Revisión: 0
	Referencia a la Norma ISO 9001:2015 8.3, 8.3.1	Página 3 de 10

4. Participantes en el diseño y seguimiento curricular del programa.

Lugar y fecha de elaboración o revisión	Participantes	Observaciones
Instituto Tecnológico de la Costa Grande.	Instituto Tecnológico de la Costa Grande, Academia de Sistemas y Computación. Ing. Irving Jesús Sánchez Campos M.C. Carlos Alberto Camacho Andrade	Reuniones de academia para el análisis de los temas establecidos en el programa de estudios.

5. Competencia(s) a desarrollar.

Competencia(s) específica(s) de la asignatura
El estudiante analiza y aplica las técnicas de extracción, transformación y carga de datos (ETL) en diferentes tipos de bases de datos y conjuntos de datos con el fin de estructurar y procesar información para el análisis. El estudiante utiliza herramientas de extracción de datos, como WEB ScrapPing y manejo de bases de datos NoSQL, para recolectar y organizar grandes volúmenes de datos provenientes de diversas fuentes. El estudiante implementa soluciones basadas en lenguajes de programación y herramientas como Python, MongoDB y otras tecnologías relacionadas para la manipulación y análisis de datos de manera eficiente y escalable.

6. Competencias previas.

El estudiante comprende y utiliza conceptos fundamentales de algoritmos y estructuras de datos, lo cual le permite organizar y manipular información de manera eficiente para tareas de procesamiento de datos. El estudiante aplica los principios básicos de bases de datos relacionales, incluyendo la creación, consulta y mantenimiento de bases de datos estructuradas, lo que le permite

	Nombre del documento: Formato de Programa de Estudio de asignatura de Especialidad	Código: TecNM-AC-PO-007-02
		Revisión: O
	Referencia a la Norma ISO 9001:2015 8.3, 8.3.1	Página 4 de 10

integrar estos conocimientos con bases de datos NoSQL y herramientas de extracción de datos.

El estudiante maneja lenguajes de programación como Python o Java, que son necesarios para implementar algoritmos de extracción y análisis de datos.

7. Temario.

No.	Temas	Subtemas
1	Introducción a la extracción de datos.	1.1 Proceso de extracción transformación y carga de datos. 1.2 Importancia de la generación de fuentes de datos. 1.2.1 Bases de datos y bases de conocimiento. 1.2.2 Conocimiento, información y datos. 1.2.3 Descubrimiento de datos. 1.3 Web Scraping y Web Crawling.
2	Fuentes y herramientas de extracción de datos.	2.1 Identificación de fuentes. 2.1.1 KAGGLE. 2.1.2 CONACYT. 2.1.3 OURWORLDINDATA. 2.1.4 INEGI. 2.2 Herramientas de extracción de datos. 2.2.1 Chrono download manager. 2.2.2 Fórmula Import HTML. 2.2.3 TABLE CAPTURE. 2.2.4 SCRAPERWIKI. 2.2.5 Tabula. 2.2.6 Import.io. 2.2.7 BeautifulSoup.
3	Extracción de datos.	3.1 Fases del ETL. 3.1.1 Extracción. 3.1.2 Transformación. 3.1.3 Carga. 3.2 Extracción de datos de Twitter. 3.3 Extracción de imágenes de Twitter.
4	Almacenamiento en bases de datos No SQL.	4.1 Bases de datos NoSQL. 4.2 Tipos de bases de datos NoSQL. 4.2.1 Bases de datos clave-valor. 4.2.2 Bases de datos documentales. 4.2.3 Bases de datos orientadas a

	Nombre del documento: Formato de		Código: TecNM-AC-PO-007-02
	Programa de Estudio de asignatura de Especialidad		Revisión: 0
	Referencia a la Norma ISO 9001:2015	8.3, 8.3.1	Página 5 de 10

		grafos. 4.2.4 Bases de datos orientadas a objetos. 4.3 Introducción a MongoDB Atlas (uso web). 4.3.1 Características principales de MongoDB Atlas. 4.3.2 Beneficios del uso de servicios en la nube. 4.4 Creación de una cuenta en MongoDB Atlas. 4.4.1 Registro en MongoDB Atlas. 4.4.2 Configuración inicial del clúster. 4.5 Puesta en marcha del servicio de MongoDB Atlas. 4.5.1 Conexión al clúster desde la web y aplicaciones externas. 4.5.2 Generación de credenciales de acceso seguro. 4.6 Gestión de bases de datos en MongoDB Atlas. 4.6.1 Crear una base de datos y una colección desde la web. 4.6.2 Agregar documentos mediante la interfaz web. 4.6.3 Administrar documentos: edición y eliminación. 4.7 Uso de MongoDB con herramientas externas. 4.7.1 Conexión mediante clientes como Python.
5	Intercambio de datos No SQL.	5.1 Importar datos hacia MongoDB Atlas desde la interfaz web. 5.1.1 Importación de archivos mediante la interfaz web. 5.1.2 Tipos de archivos compatibles (.csv, .json). 5.2 Exportar datos desde MongoDB Atlas 5.2.1 Exportación de colecciones en formato .csv. 5.2.2 Exportación de colecciones en formato .json. 5.3 Manipulación de datos en MongoDB Atlas mediante Python. 5.3.1 Conexión a MongoDB Atlas desde Python. 5.3.2 Agregar documentos desde un

	Nombre del documento: Formato de		Código: TecNM-AC-PO-007-02
	Programa de Estudio de asignatura de Especialidad		Revisión: 0
	Referencia a la Norma ISO 9001:2015	8.3, 8.3.1	Página 6 de 10

		script de Python. 5.3.3 Actualizar documentos desde un script de Python. 5.3.4 Eliminación de documentos desde un script de Python.
--	--	--

8. Actividades de aprendizaje de los temas.

Tema 1: Introducción a la extracción de datos.	
Competencias	Actividades de aprendizaje
Específica(s): Aplicar conceptos básicos de extracción de datos y conocer las herramientas necesarias para su obtención, procesamiento y análisis, comprendiendo la importancia de las fuentes de datos y las prácticas éticas en el uso de técnicas de web scraping y crawling. Genéricas: • Capacidad de aprendizaje autónomo. • Pensamiento crítico y analítico. • Habilidad para resolver problemas.	• Diseñar un esquema visual del proceso de Extracción, Transformación y Carga (ETL), explicando cada una de las fases con ejemplos sencillos. • Investigar una fuente de datos abierta (p. ej., datos gubernamentales), explicando su uso potencial y el tipo de información que contiene. • Realizar una práctica guiada en la que extraerán información de una página web de ejemplo utilizando la herramienta BeautifulSoup en Python, respetando los términos de servicio.
Tema 2: Fuentes y herramientas de extracción de datos.	
Competencias	Actividades de aprendizaje
Específica(s): Identificar fuentes de datos confiables y utilizar herramientas de extracción de datos adecuadas para recopilar y procesar información de diferentes fuentes, comprendiendo las características y el uso de cada herramienta en función de los objetivos del análisis. Genéricas: • Capacidad para buscar, seleccionar y utilizar información. • Pensamiento analítico y crítico.	• Investigar las plataformas Kaggle, CONACYT, OURWORLDINDATA, INEGI y seleccionarán una fuente de datos para un tema de interés. Deberán presentar una breve descripción del conjunto de datos y su posible utilidad en un proyecto. • Crear una tabla comparativa entre las fuentes (Kaggle, CONACYT, OURWORLDINDATA, INEGI) indicando características como tipo de datos, frecuencia de actualización, y accesibilidad.

	Nombre del documento: Formato de Programa de Estudio de asignatura de Especialidad	Código: TecNM-AC-PO-007-02
		Revisión: 0
	Referencia a la Norma ISO 9001:2015 8.3, 8.3.1	Página 7 de 10

Habilidad en el uso de tecnologías de la información.	- Extraer datos de una página web utilizando BeautifulSoup en Python. La actividad será guiada para evitar problemas legales o éticos, trabajando en sitios de datos abiertos. - Cada estudiante investigará sobre Tabula y Import.io y realizará una breve presentación explicando el tipo de datos que se pueden extraer con estas herramientas y sus aplicaciones prácticas.
---	--

Tema 3: Extracción de datos.

Competencias	Actividades de aprendizaje
Específica(s): Aplicar el proceso ETL y técnicas de extracción de datos de redes sociales, comprendiendo el flujo de extracción, transformación y carga, así como la obtención de datos e imágenes de plataformas como Twitter de manera ética y eficaz. Genéricas: - Habilidad para resolver problemas. - Pensamiento analítico. - Uso de tecnologías de la información.	- Investigación sobre cómo diferentes empresas utilizan el proceso ETL para sus operaciones de datos. Los estudiantes presentarán un caso que incluya ejemplos de extracción, transformación y carga de datos en una empresa real. - Los estudiantes crearán un script en Python para realizar un proceso ETL sencillo. Extraerán datos de un archivo CSV, aplicarán una transformación básica (por ejemplo, limpieza de datos) y luego cargarán el resultado en un nuevo archivo. - Los estudiantes aprenderán a configurar el acceso a la API de Twitter (usando Tweepy o similar en Python), realizar búsquedas de tweets y extraer información básica como el texto, el autor y la fecha.

Tema 4: Almacenamiento en bases de datos No SQL.

Competencias	Actividades de aprendizaje
Específica(s): Configurar y gestionar bases de datos No SQL, especialmente MongoDB, comprendiendo sus tipos y aplicando técnicas básicas de instalación,	Investigar y presentarán las diferencias entre bases de datos SQL y No SQL, resaltando ventajas y desventajas de cada tipo.

	Nombre del documento: Formato de		Código: TecNM-AC-PO-007-02
	Programa de Estudio de asignatura de Especialidad		Revisión: 0
	Referencia a la Norma ISO 9001:2015	8.3, 8.3.1	Página 8 de 10

puesta en marcha y manipulación de datos en colecciones. Genéricas: • Uso de tecnologías de la información. • Capacidad de aprendizaje autónomo. • Habilidad para resolver problemas.	• Crear un mapa conceptual que represente los diferentes tipos de bases de datos No SQL (clave-valor, documentales, orientadas a grafos y orientadas a objetos) y ejemplos de cada uno. • Guiar a los estudiantes en el proceso de uso de MongoDB atlas (web). • Los estudiantes crearán una base de datos en MongoDB, así como operaciones sencillas (agregar, editar y eliminar).
Tema 5: Intercambio de datos No SQL	
Competencias	Actividades de aprendizaje
Específica(s): Manejar la importación, exportación y manipulación de datos en bases de datos No SQL, en especial MongoDB, utilizando herramientas como Compass y Python para la gestión de archivos en formatos CSV y JSON. Genéricas: • Uso de tecnologías de la información. • Habilidad para resolver problemas. • Capacidad de aprendizaje autónomo.	• Realizar una práctica donde importarán archivos CSV y JSON a MongoDB. • Los estudiantes exportarán datos a archivos CSV y JSON a MongoDB, observando el formato de los datos en el archivo resultante. • Utilizarán un script en Python (con la biblioteca PyMongo) para agregar datos a una colección en MongoDB. • Implementarán un script en Python que actualice registros específicos en una colección de MongoDB, basado en condiciones establecidas.

9. Práctica(s).

• Importa un archivo CSV a MongoDB usando Compass y explora los datos importados. Examina la estructura y el contenido de los documentos para entender cómo se representan en MongoDB. • Usa Compass para exportar una colección de MongoDB a formatos CSV y JSON. Compara ambos archivos para observar las diferencias en la estructura de datos.

	Nombre del documento: Formato de		Código: TecNM-AC-PO-007-02
	Programa de Estudio de asignatura de Especialidad		Revisión: 0
	Referencia a la Norma ISO 9001:2015	8.3, 8.3.1	Página 9 de 10

- Importa un archivo JSON a MongoDB usando la consola. Luego, realiza una exportación de esa misma colección en formato CSV, verificando que la estructura se mantenga.
- Crea un script en Python que use PyMongo para insertar documentos en una colección de MongoDB. Luego, añade otra función al script para actualizar un campo específico en esos documentos.
- Realiza operaciones CRUD (Crear, Leer, Actualizar, Borrar) sobre una colección en MongoDB usando Compass y la consola. Asegúrate de documentar cada operación y el resultado obtenido.

10. Proyecto de asignatura.

- **Fundamentación:** Los estudiantes realizarán un marco referencial en el que investigarán y analizarán los conceptos y teorías sobre bases de datos No SQL y su relevancia en el contexto actual de manejo de datos masivos. Deberán realizar un diagnóstico sobre las necesidades de la organización (real o ficticia) que utilizará el sistema, enfocándose en los beneficios de las bases de datos No SQL frente a las SQL tradicionales en este contexto.
- **Planeación:** Con base en el diagnóstico realizado, los estudiantes diseñarán el sistema de gestión de datos, planificando las actividades, recursos y cronograma de trabajo. Este diseño debe incluir el tipo de datos que se manejarán, el modelo de almacenamiento en MongoDB, y los procesos de importación y exportación de datos necesarios.

Ejecución: Los estudiantes implementarán el sistema de gestión de datos, creando y configurando las bases de datos en MongoDB, desarrollando scripts en Python para la manipulación de datos, y configurando procesos de importación y exportación. Durante esta fase, se deberán realizar pruebas para asegurar el correcto funcionamiento del sistema.
- **Evaluación:** Al finalizar el proyecto, los estudiantes evaluarán el sistema implementado, analizando los logros alcanzados y los aspectos a mejorar. Además, realizarán una reflexión sobre el desarrollo de sus competencias técnicas y genéricas. La evaluación incluirá tanto el funcionamiento técnico del sistema como el impacto potencial en la organización ficticia o real.

	Nombre del documento: Formato de		Código: TecNM-AC-PO-007-02
	Programa de Estudio de asignatura de Especialidad		Revisión: 0
	Referencia a la Norma ISO 9001:2015	8.3, 8.3.1	Página 10 de 10

11. Evaluación por competencias.

Para evaluar este proyecto, se recomienda solicitar evidencias como el reporte del marco referencial, diagramas de flujo del sistema, reportes de prácticas de configuración de MongoDB, código de scripts en Python, bitácoras de pruebas y una presentación final del sistema implementado.

Para verificar el logro de competencias, se sugiere utilizar listas de cotejo para la configuración de MongoDB, listas de verificación para la funcionalidad de los scripts, matrices de valoración para el diseño y diagramas, guías de observación para la participación en prácticas y rúbricas para evaluar los reportes y la presentación final, enfocándose en claridad, precisión y calidad técnica.

12. Fuentes de información.

1. Chodorow, K. (2013) MongoDB: The Definitive Guide. USA: O'Reilly Media. Disponible en e-libro.
2. Marz, N. (2013) Big Data: Principles and Best Practices of Scalable Real-Time Data Systems. USA: Manning Publications. Disponible en e-libro.
3. Sadalage, P. J. y Fowler, M. (2012) NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. USA: Addison-Wesley. Disponible en e-libro.
4. Google Scholar. (2024) No SQL databases y MongoDB. Disponible en <https://scholar.google.com/>
5. Khan Academy. (2024) Bases de datos y manejo de datos. Disponible en <https://www.khanacademy.org/>
6. MongoDB University. (2024) Introduction to MongoDB y Data Modeling. Disponible en <https://university.mongodb.com/>
7. GitHub. (2024) Repositorios de proyectos en MongoDB y Python. Disponible en <https://github.com/>