



Introducción a Big Data

1. Datos Generales de la asignatura

Nombre de la asignatura:	<i>Introducción a Big Data</i>
Clave de la asignatura:	IAC-2206
SATCA¹:	2-2-4
Carrera:	Ingeniería en Sistemas Computacionales

2. Presentación

Caracterización de la asignatura
Con esta asignatura el estudiante conocerá la generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos.
Como aportación al perfil de egresado se tiene lo siguiente:
<ul style="list-style-type: none">• Identificar y resolver problemas relacionados con el paradigma Big Data• Clasificar y almacenar los diferentes tipos de datos de este paradigma• Realizar análisis de datos de diferentes fuentes (Web, redes sociales, M2M, Cloud, entre otras)• Poseer las habilidades metodológicas de investigación para sentar las bases de conocimiento para futuros Científicos de Datos
La asignatura de Introducción a Big Data, presenta al estudiante el panorama actual del manejo de datos en un mundo globalizado, donde la información que se genera se mide en zettabytes (1 billón de GB). Esto genera la necesidad de contar con Científicos de Datos que respondan a la demanda actual de almacenar, analizar y presentar la información creciente, de manera que las empresas realicen negocios a partir de esta información.
Esta asignatura presenta en el tema uno el panorama que comprende el concepto y características de Big Data, el cual no sólo se refiere a grandes volúmenes de datos, si no que abarca las diferentes fuentes de datos estructurados, no estructurados y semi estructurados que requieren ser almacenados en especializadas bases de datos.
En el tema dos, se tratará una introducción sobre la ciencia de los datos, de tal manera que se conozcan diferentes entornos de esta ciencia, así como algunas tecnologías adecuadas para el desarrollo de aplicaciones pertinentes.

¹ Sistema de Asignación y Transferencia de Créditos Académicos



En el tema tres, se conocerán diferentes herramientas para hacer el análisis, tratamiento y curación de los datos, aplicando tecnologías analíticas.

El tema cuatro se centra en conocer diferentes plataformas de Big Data, mediante las cuales se pueda hacer el análisis y manejo de grandes volúmenes de datos.

Asignaturas con las que se relaciona la materia:

Administración de Bases de Datos, Probabilidad y Estadística, Gestión de Proyectos de Software, Programación Web, Sistemas Embebidos.

Como competencia específica de esta asignatura, el estudiante podrá identificar y analizar grandes volúmenes de información, para su tratamiento y automatización, de tal manera que pueda ser utilizada para la toma de decisiones.

Intención didáctica

En el tema uno, se investigarán, comprenderán y asimilarán los conceptos básicos para definir el paradigma Big Data. Es importante, ver todo el panorama. Después de investigar la definición y las características, se debe puntualizar en los diferentes tipos de datos, tomando en cuenta que sólo una parte de los datos son estructurados. Las fuentes de donde provienen todos estos datos son muy variadas, se deben investigar tanto conceptos como ejemplos de la Web, el Internet de las cosas, Biométrica, teléfonos inteligentes, los cuales son fuentes directas de Big Data.

En el tema dos se conocerán diferentes entornos de las ciencias de datos, con la finalidad de que se puedan distinguir las aplicaciones que se pueden hacer con éstos, así como, las diferentes tecnologías que pueden emplearse para la ciencia de datos.

La unidad tres se enfoca al análisis de datos, que comprende diversas técnicas para el tratamiento, curación y manejo de la información. Se propone investigar qué software especializado existe según el tipo de información que se desea obtener y estudiar ejemplos específicos para cada tópico.

En el tema cuatro, el alumno desarrollará una serie de prácticas donde se estudien diferentes plataformas de análisis de datos, como Map Reduce, Hadoop, Spark, Hive e Impala. Se sugiere utilizar Hadoop, con Map Reduce.

Competencias Genéricas

- Capacidad de análisis y síntesis.
- Conocimientos básicos de la carrera.
- Comunicación oral y escrita.
- Habilidades básicas de manejo de la computadora.
- Habilidad para buscar y analizar información proveniente de fuentes diversas.



3. Participantes en el diseño y seguimiento curricular del programa

Lugar y fecha de elaboración o revisión	Participantes	Observaciones
Cd. Guzmán Jalisco, Enero 2022.	Instituto Tecnológico de Cd. Guzmán.	Diseño de la especialidad para el plan de estudios de la carrera de Ingeniería en Sistemas Computacionales ISIC-2010-224.

4. Competencia(s) a desarrollar

Competencia(s) específica(s) de la asignatura
Aplica herramientas y plataformas al análisis de grandes volúmenes de datos para la resolución de problemas del entorno.

5. Competencias previas

- Utiliza técnicas de modelado para la solución de problemas.
- Comprende los conceptos de las bases de datos.
- Construye modelos de bases de datos.
- Domina el lenguaje SQL.
- Diseña, crea y administra esquemas de bases de datos relacionales bajo estándares.
- Conoce distintas plataformas y SGBD.
- Aplica las actividades que involucra cada una de las etapas del ciclo de vida de un proyecto de software.
- Analiza, diseña, desarrolla e implementa un proyecto de software.
- Conoce e identifica modelos de arquitecturas de cómputo.
- Identifica y aplica conceptos fundamentales de las telecomunicaciones, para aplicarlo a redes computacionales.
- Selecciona, conoce y usa adecuadamente los diferentes sistemas operativos para lograr un uso más eficiente, así como diferenciar y aplicar las técnicas de manejo de recursos para el diseño, organización, utilización y optimización de los sistemas operativos.
- Conoce y sabe usar técnicas y/o herramientas de administración de los sistemas operativos para la optimización de recursos existentes.
- Aplica un lenguaje orientado a objetos para la solución de problemas.
- Crea y manipula bases de datos utilizando distintos Gestores de Bases de Datos considerando elementos de integridad y seguridad para el tratamiento de la información en distintas plataformas.



6. Temario

No.	Temas	Subtemas
1	Introducción a Big Data	1.1 Conceptos y características 1.2 Tipos de datos (estructurados, no estructurados y semiestructurados) 1.3 Componentes de un sistema de Big Data 1.4 Almacenamiento y modelos de datos 1.5 Modelos de programación
2	Introducción a la ciencia de los datos	2.1 Entornos de ciencia de datos 2.2 Tecnologías para la ciencia de datos
3	Herramientas de análisis de datos	3.1 Tratamiento y curación de los datos 3.2 Aplicaciones y tecnologías analíticas
4	Tecnologías de Big Data	4.1 Map Reduce 4.2 Hadoop 4.3 Spark 4.4 Hive e Impala

7. Actividades de aprendizaje de los temas

Nombre de tema	
Introducción a Big Data	
Competencias	Actividades de aprendizaje
Específica(s): Comprende la importancia de Big Data, así como sus características y arquitectura. Genéricas: Competencias instrumentales: <ul style="list-style-type: none">• Capacidad de análisis y síntesis• Capacidad de organizar y planificar• Comunicación oral y escrita en su propia lengua• Capacidad de lectura en una segunda lengua	<ul style="list-style-type: none">• Comprender la definición de Big Data y conocer su importancia.• Conocer las características y la arquitectura básica de Big Data.• Identificar los tipos de datos: estructurados, no estructurados y semi estructurados.• Conocer los componentes y funciones específicas de Big Data.• Investigar los modelos de almacenamiento de datos y de programación.



<ul style="list-style-type: none">• Habilidades de gestión de información (habilidad para buscar y analizar información proveniente de fuentes diversas)• Solución de problemas• Toma de decisiones	
Nombre de tema	
Introducción a la ciencia de los datos	
Competencias	Actividades de aprendizaje
<p>Específica(s):</p> <p>Conoce los entornos y las tecnologías apropiadas para la ciencia de datos.</p> <p>Genéricas:</p> <ul style="list-style-type: none">• Capacidad de análisis y síntesis• Capacidad de organizar y planificar• Comunicación oral y escrita en su propia lengua• Capacidad de lectura en una segunda lengua• Habilidades de gestión de información (habilidad para buscar y analizar información proveniente de fuentes diversas)• Solución de problemas• Toma de decisiones	<ul style="list-style-type: none">• Investigar para conocer los diferentes entornos que existen para la ciencia de datos.• Investigar diversas aplicaciones y tecnologías para la ciencia de datos, para seleccionar las más adecuadas para desarrollar un pequeño proyecto.
Nombre de tema	
Herramienta de análisis de datos	
Competencias	Actividades de aprendizaje
<p>Específica(s):</p> <p>Conoce las herramientas de análisis de datos para su tratamiento y curación.</p> <p>Genéricas:</p> <ul style="list-style-type: none">• Capacidad de análisis y síntesis• Capacidad de organizar y planificar	<ul style="list-style-type: none">• Hacer una práctica que implique el tratamiento y curación de los datos.• Investigar diversas aplicaciones y tecnologías analíticas, para poder seleccionar las más adecuadas para desarrollar un pequeño proyecto.



<ul style="list-style-type: none">• Comunicación oral y escrita en su propia lengua• Capacidad de lectura en una segunda lengua• Habilidades de gestión de información (habilidad para buscar y analizar información proveniente de fuentes diversas)• Solución de problemas• Toma de decisiones	
Nombre de tema	
Tecnologías de Big Data	
Competencias	Actividades de aprendizaje
<p>Específica(s):</p> <p>Conoce diferentes tecnologías de Big Data que permitan hacer el análisis de grandes volúmenes de datos.</p> <p>Genéricas:</p> <ul style="list-style-type: none">• Capacidad de análisis y síntesis• Capacidad de organizar y planificar• Comunicación oral y escrita en su propia lengua• Capacidad de lectura en una segunda lengua• Habilidades de gestión de información (habilidad para buscar y analizar información proveniente de fuentes diversas)• Solución de problemas• Toma de decisiones	<ul style="list-style-type: none">• Desarrollar prácticas que implementen cada una de las siguientes tecnologías de Big Data:<ul style="list-style-type: none">○ Map Reduce○ Hadoop○ Spark○ Hive e Impala

8. Práctica(s)

- Instalar un entorno virtual pre-configurado (Cloudera o HortonWorks) que contenga las tecnologías para la ciencia de datos: Hadoop, Spark, Hive e Impala. Se harán ejercicios relacionados al manejo y administración de archivos con HDFS.
- Desarrollar un programa utilizando el Map Reduce de Hadoop/Java. Se llevarán a cabo experimentos y un reporte de la ejecución versus una versión secuencial.



- Desarrollar una serie de ejercicios de programación utilizando Python Spark. Realizar una comparativa en desempeño de una versión del programa en Python Spark versus el de Hadoop/Java.
- Realizar ejercicios de consultas de datos a través de Apache Hive e Impala.
- Desarrollar un proyecto final del curso donde combine el uso de Python Spark con la Consulta de datos (con Hive e Impala).

9. Proyecto de asignatura

El objetivo del proyecto que planteé el docente que imparte esta asignatura, es demostrar el desarrollo y alcance de la(s) competencia(s) de la asignatura, considerando las siguientes fases:

- **Fundamentación:** marco referencial (teórico, conceptual, contextual, legal) en el cual se fundamenta el proyecto de acuerdo con un diagnóstico realizado, mismo que permite a los estudiantes lograr la comprensión de la realidad o situación objeto de estudio para definir un proceso de intervención o hacer el diseño de un modelo.
- **Planeación:** con base en el diagnóstico en esta fase se realiza el diseño del proyecto por parte de los estudiantes con asesoría del docente; implica planificar un proceso: de intervención empresarial, social o comunitario, el diseño de un modelo, entre otros, según el tipo de proyecto, las actividades a realizar los recursos requeridos y el cronograma de trabajo.
- **Ejecución:** consiste en el desarrollo de la planeación del proyecto realizado por parte de los estudiantes con asesoría del docente, es decir en la intervención (social, empresarial), o construcción del modelo propuesto según el tipo de proyecto, es la fase de mayor duración que implica el desempeño de las competencias genéricas y específicas a desarrollar.
- **Evaluación:** es la fase final que aplica un juicio de valor en el contexto laboral-profesión, social e investigativo, ésta se debe realizar a través del reconocimiento de logros y aspectos a mejorar se estará promoviendo el concepto de “evaluación para la mejora continua”, la metacognición, el desarrollo del pensamiento crítico y reflexivo en los estudiantes.

10. Evaluación por competencias

Son las técnicas, instrumentos y herramientas sugeridas para constatar los desempeños académicos de las actividades de aprendizaje.

Se recomienda evaluar los siguientes aspectos:

- Asistencia.
- Investigación.
- Participación en el aula.
- Cuestionarios.
- Prácticas de laboratorio.



- Trabajo en equipo.
- Proyecto de aplicación.

11. Fuentes de información

Bibliografía básica:

- Butch, Q. (2018). Next-Generation Big Data, a practical guide to Apache Kudu, Impala and Spark. Primera edición. USA: Apress.
- Capriolo, E., Wampler, D. & Rutherglen, J. (2012). Programming Hive. USA: O'Reilly Media.
- Chambers, B. & Zaharia, M. (2018). Spark the definitive guide. USA: O'Reilly Media, Inc.
- García, A. M. (2017). Big Data gestión y explotación de grandes volúmenes de datos. México: Editorial UOC, S. L.
- Joyanes_Aguilar, L. (2013). Big Data, análisis de grandes volúmenes de datos en organizaciones. (pág. 428). México: Alfaomega.
- Lazyprogrammer. (2016). Big Data, MapReduce, Hadoop and Spark with Python. Disponible en <https://lazyprogrammer.me>.
- López M. J., Zarza G. (2017). La ingeniería del Big Data: cómo trabajar con datos. Barcelona. Editorial UOC.
- Lublinsky, B., Smith, K. T. & Yakubovich, A. (2014). Hadoop: soluciones Big Data. España: Anaya Multimedia.
- Mayer-Schönberger, V., & Cukier, K. (2013). Big data. La revolución de los datos masivos (Noema). (pág. 278). Madrid: Turner Naoma.
- Menoyo, R. D., García, L. E., García, C. A. (2021). Fundamentos de la ciencia de datos. España. Editorial Universidad de Alcalá.
- Ríos I. D., Gómez-Ullate O. D. (2019). Big Data: Conceptos, tecnologías y aplicaciones. Madrid. CSIC Editorial.
- Russell, J. (2014). Getting Started with Impala. USA: O'Reilly.

Bibliografía complementaria:

- 4C Infotech. (s.f.). Aprenda Big Data y Hadoop gratuito (App gratis).
- Angeles, C. M., Cilleros, S. D. (2020) Ciberseguridad y transformación digital. Cloud, Identidad Digital, Blockchain, Agile, Inteligencia Artificial. España. Anaya Multimedia.
- Caballero, R., Marín, E. y Rieso, A. (2018). Big Data con Phyton. RCLIA|#RC LIBROS.
- Camilo U. C., García P. M., Osorio Q. J., Navarro C. A. (2018). Ciberseguridad: un enfoque desde la ciencia de datos. Colombia: Editorial Universidad Icesi.
- Castillo Romero, J. (2019). Big Data. Manual de Especialidad Formativa. Málaga. IC Editorial.
- García, A. M. (2017). Big Data gestión y explotación de grandes volúmenes de datos. México: Editorial UOC, S. L.
- Provost, F., & Fawcett, T. (2013). Data Science for Business. (pág. 384). Sebastopol, CA: O'reilly & Associates Inc.



- Rabasa, A., & Brebbia, C. A. (2016). BIG DATA. (Vol. 1, pág. 300). NY: Wit Pr/Computational Mechanics.
- Rungta, K. (2018). Hadoop. Aprende en 1 día. Kindle Edition.
- White, T. (2015). Hadoop the definitive guide. Cuarta edición. USA: O'Reilly Media, Inc.