Datos Generales

1. Nombre de la Asignatura Visualización de Grandes Bases de Datos	2. Nivel de formación Maestría		3. Clave de la Asignatura
4. Prerrequisitos Cursos de Procesamiento y almacenamiento de Grandes Bases de Datos	5. Área de Formación Formación Básica Particular Obligatoria		6. Departamento Sistemas de Información
7. Modalidad Presencial	8. Tipo de Asignatura Curso - Taller		
9. Carga Horaria 48 horas			
Teoría: 30	Práctica: 18	Total: 48	Créditos: 7
10. Trayectoria de la asignatura			

Contenido del Programa

11. Presentación
Las disciplinas de la informática y las ciencias de la computación se encuentran inmersas en una revolución tecnológica originada de la creciente necesidad de acumular datos generados de manera masiva, lo que comúnmente se conoce con grandes bases de datos. La aplicación masiva de la inteligencia artificial, mediante el uso de la combinación de diferentes paradigmas y métodos de análisis de datos han permitido desarrollar nuevas maneras de interpretar información que apoyen la toma de decisiones y la inferencia estadística basada en modelos matemáticos. Este curso parte del núcleo académico básico del posgrado de ciencia de datos pretende proporcionar la bases teóricas y técnicas que permitan explorar e interpretar información obtenida de grandes bases de datos empleando diferentes estrategias mediante comandos en línea y herramientas en plataforma.
12.- Objetivos del programa Adquirir los conocimientos básicos necesarios para la visualización de grandes bases de datos. Desarrollar habilidades y destrezas básicas para análisis de datos y la identificación de patrones. Comprender los principios básicos de la visualización y representación de grandes bases de datos para poder facilitar su interpretación. Inferir patrones, distribuciones y tendencias partir de grandes bases de datos. Desarrollar competencias básicas de acceso y velocidad con *Hadoop* y *Spark*. Entender los pasos básicos para seleccionar y desplegar datos obtenidos de grandes repositorios.

Objetivo General

Desarrollar las habilidades y destrezas básicas para la visualización y el análisis de grandes bases de datos.

13.-Contenido

Contenido temático

1. Introducción a los métodos de visualización de grandes bases de datos
2. Principales *frameworks* de trabajo en grandes bases de datos
3. Extendiendo el software *stack* de grandes bases de datos
4. *Framework Apache Spark*
5. Conjuntos de datos
6. Formatos de archivos
7. *Data frames* y consultas
8. Receptores, transformaciones y operaciones de salida
9. Utilidad de herramientas para procesamiento de datos estructurados y semi-estructurados
10. *Framework Apache Hadoop*
11. Herramientas para visualización de datos
12. Herramientas para visualización de datos avanzadas

Capítulo 1 Introducción a los métodos de visualización de grandes bases de datos

Objetivos particulares del capítulo:
Conocer los antecedentes y la evolución histórica de los sistemas de grandes bases de datos
Definir que son los algoritmos y los fundamentos de los métodos de paralelización
Aprender los fundamentos de aprendizaje automático.

Desarrollo

1.1 Introducción a las grandes bases de datos
1.1.1 Antecedentes y contextualización
1.1.2 EL nuevo paradigma de las grandes bases de datos
1.1.3 Utilidad ¿dónde encontramos grandes bases de datos?
1.1.4 Datos, información y conocimiento.
1.1.5 Ejemplos de escenarios de grandes bases de datos
1.2 Algoritmos, paralelización y grandes bases de datos
1.2.1 La problemática del procesado secuencial y el volumen de datos
1.2.2 ¿Qué es un algoritmo?
1.2.3 Eficiencia de implementación de algoritmos.
1.3 Introducción al aprendizaje automático
1.3.1 Tipología de métodos
1.3.2 Tipología de tareas
1.3.3 Fases de un proyecto de aprendizaje automático
1.3.4 Redes neuronales y aprendizaje profundo (*Deep learning*)

Capítulo 2 Principales *frameworks* de trabajo en grandes bases de datos

Objetivo: Identificar las ventajas y desventajas de los ambientes de trabajo *Hadoop* y *Spark* para grandes bases de datos.

Desarrollo:
Lectura y discusión del artículo *Actualidad e importancia de la implementación de Big Data utilizando las herramientas Hadoop y Spark*. Lámpsakos, (19), 67-72. Autores: Gil Restrepo, Gustavo Andrés - Montoya Suarez, Lina. ISSN: 21454080 Editorial: Universidad Católica Luis Amigó Año de Edición: 2018.

Capítulo 3 *Framework Apache Spark*

Objetivos particulares del capítulo:
Aprender las generalidades sobre el entorno *Apache spark*
Valorar la flexibilidad de *spark para la* interacción con lenguajes de programación como *Python* y *R*
Conocer otras alternativas al entorno *spark*.

Desarrollo:

3.1 *Spark* y *Python* 3.2 *Spark* y sus alternativas
Capítulo 4 Instalación de *Apache Spark*
Objetivos particulares del capítulo: Aprender a descargar *Apache spark* Usar llenas de comando en *Python* en un entono *spark* Conocer los conceptos esenciales y otras alternativas al entorno *spark*. Desarrollar aplicaciones autocontenidas y Configurar *Spark*
Desarrollo: 4.1 Descargar Apache *Spark* 4.2 Introducción al *shell* de *Python* 4.3 Conceptos esenciales de *Spark* 4.4 Aplicaciones autocontenidas 4.5 Configurando *Spark*
Capítulo 5 Conjuntos de datos
Objetivos particulares del capítulo: Identificar conjuntos de datos resilientes y distribuidos Definir variables compartidas
Desarrollo: 5.1 Conjuntos de datos resilientes y distribuidos 5.2. Variables compartidas
Capítulo 6 Formatos de archivos
Objetivos particulares del capítulo: Conocer los formatos de archivos Aprender a formatear archivos Conocer las generalidades de bases de datos en el entorno *Apache spark*
Desarrollo: 6.1 Formatos de archivos 60.2 Bases de datos
Capítulo 7 *Data frames* y consultas
Objetivos particulares del capítulo: Conocer las generalidades sobre *Data frames* Aprender a realizar consultas *SQL* en el ambiente *Apache spark*

Desarrollo: 7.1 *Data frames* 7.2 Consultas *SQL*
Capítulo 8 Receptores, transformaciones y operaciones de salida
Objetivos particulares del capítulo: Aprender sobre receptores, transformaciones y operaciones de salida con *DStreams* Trabajar con *DataFrames* y operaciones *SQL* con *SparkStreaming*
Desarrollo: 8.1 Un ejemplo sencillo 8.2 Receptores 8.3 Transformaciones 8.4 Operaciones de salida con *DStreams* 8.5 *DataFrames* y operaciones *SQL* con *SparkStreaming*
Capítulo 9 Utilidad de herramientas para procesamiento de datos estructurados y semiestructurados
Objetivos particulares del capítulo: Aprender sobre el uso del módulo *MLlib* y sus *f*uncionalidades
Desarrollo: 9.1 El módulo *MLlib* 9.2 Uso de *MLlib* 9.3 Funcionalidades de *MLlib*
Capítulo 10 *Hadoop*
Objetivos particulares del capítulo: Configurar el entorno en *Apache Hadoop* Describir las generalidades del Sistema *HDFS* Conocer las operaciones básicas en *HDFS* Referir comandos Emplear *MapReduce* y *Streaming* Conocer las funcionalidades sobre varios nodos de clúster
Desarrollo: 10.1 Grandes soluciones de datos para grandes datos generales 10.2 Configuración entorno en *Apache Hadoop* 10.3 Descripción general del Sistema de Archivos Distribuidos *Hadoop (HDFS*) 10.4 Operaciones en *HDFS* 10.5 Referencia de comandos

10.6 *MapReduce*
10.7 *Streaming*
10.8 Varios nodos de clúster

Capítulo 11 Herramientas para visualización de datos

Objetivos particulares del capítulo:
Conocer algunas de las principales herramientas básicas para visualizar productos de grandes bases de datos

Desarrollo:
11.1 Herramientas de depuración y filtrado
11.2 Herramientas de análisis y exploración
11.3. Herramientas para gráficos de datos
11.4 Herramientas para visualización de datos
11.5 Herramientas de visualización de mapas
11.6 Herramientas para crear infografías
11.7 Google Charts

Capítulo 12 Herramientas para visualización de datos avanzadas

Objetivos particulares del capítulo:
Conocer algunas de herramientas avanzadas para visualizar productos de grandes bases de datos y que requieren conocimientos en *Java*, *Python* y *R* entre otros lenguajes de programación

12.1 *D3.js*
12.2 *Ember Charts*
12.3 *NVD3*
12.4 *FusionCharts*
12.5 *Highcharts*
12.6 *Chart.js*
12.7 *leaflet*
12.8 *Chartis.js*
12.9 *n3-charts*
12.10 *Sigma JS*
12.11 *ggplot* en *R*

14. Actividades Prácticas
El curso está diseñado con una visión de aprender haciendo, basado en un enfoque de enseñanza basada en problemas, tomando en cuenta que el grupo esta integrados con alumnos con diferentes habilidades y destrezas derivadas de su formación profesional diversa.
Se estimula el proceso de aprendizaje mediante las lecturas del texto de referencia fuera del aula y la resolución de tareas en el aula virtual del curso, complementándose con ejercicios prácticos (manos a la obra) en entornos *Spark y Hadoop,* así como diversas paqueterías para visualización durante sesiones presenciales parte del taller.

15.- Metodología
El curso taller tiene un enfoque práctico para facilitar el proceso de enseñanza aprendizaje,

complementando las lecturas y actividades del libro de texto, con ejercicios prácticos durante las sesiones practicas semanales.

16. Evaluación

Reportes de actividades y ejercicios semanales en aula virtual 60%
Tareas 20%
Actividades presenciales en el aula durante talleres presenciales 20%

17.- Bibliografía

BIG DATA: ANÁLISIS DE DATOS EN ENTORNOS MASIVOS. Autores: Nin Guerrero, Jordi - Casas Roma, Jordi - Julbe López, Francesc. ISBN: 9788491804727, 9788491804734 Editorial: Editorial UOC, Año de Edición: 2019

INTRODUCCIÓN A *APACHE SPARK*: PARA EMPEZAR A PROGRAMAR EL BIG DATA. Autores: Macías, Mario - Mauro Gómez ISBN: 9788491160373, 9788491160458, Editorial: Editorial UOC, Año de Edición: 2015.

VISUALIZACIÓN DE LA INFORMACIÓN: DE LOS DATOS AL CONOCIMIENTO. Autores: Ignasi Alcalde ISBN: 9788490647523, 9788497884921, Editorial UOC Año de Edición: 2015.

Otros materiales

BIG DATA CON *PYTHON* - Recolección, almacenamiento y proceso. AUTOR (ES): CABALLERO ROLDÁN, Rafael; MARTÍN MARTÍN, Enrique; RIESCO RODRÍGUEZ, AdríanISBN: 9786075383712. NÚMERO DE EDICIÓN: 1, CATIDAD DE PÁGINAS: 284. EDITORIAL: Alfaomega, RC Libros.

R FOR DATA SCIENCE. Autores: Hadley Wickham, Garrett Grolemund. Released December 2016. Publisher(s): O'Reilly Media, Inc. ISBN: 9781491910382Versión en español: https://es.r4ds.hadley.nz/

BIG DATA ANALYTICS WITH R, Autores: Simon Walkowiak, Released July 2016, Publisher(s): Packt Publishing, ISBN: 9781786466457

Herramientas de visualización

Tableau	https://www.tableau.com/
Infogram	https://infogram.com/
ChartBlocks	https://www.chartblocks.com/en/
Datawrapper	https://www.datawrapper.de/
Ploty	https://plotly.com/
RAW	https://rawgraphs.io/
Visually	https://visual.ly/

Herramientas de visualización avanzadas

D3.js	https://d3js.org/
Ember Charts	https://opensource.addepar.com/ember-charts/#/overview

NVD3	http://nvd3.org/
Google Charts	https://developers.google.com/chart/
FusionCharts	https://www.fusioncharts.com/
Highcharts	https://www.highcharts.com/
Chart.js	https://www.chartjs.org/
leaflet	https://leafletjs.com/
Chartis.js	https://gionkunz.github.io/chartist-js/
n3-charts	https://github.com/n3-charts
Sigma JS	http://sigmajs.org/
polymaps	http://polymaps.org/
Processing.js	https://processing.org/

18.- Perfil del profesor

El profesor contará con grado de Maestro o Doctor en alguna de las LGAC del Programa, es especial que cuente con conocimientos sólidos en el manejo de grandes volúmenes de datos.

19.- Nombre de los profesores que imparten la materia

Dr. Raúl Cuauhtémoc Baptista Rosas

20.- Lugar y fecha de su aprobación (incluyendo la última actualización)

Zapopan, Jal. 13 de enero de 2020

21.- Instancias que aprobaron el programa (Junta Académica y/ó Coordinación del programa)

Junta Académica y Coordinación de la Maestría en Ciencia de los Datos, y el Departamento de Sistemas de Información.