Datos Generales

1. Nombre de la Asignatura	2. Nivel de formación	3. Clave de la Asignatura
Procesamiento de Grandes Bases de Datos	Maestría	IH595
4. Prerrequisitos	5. Área de Formación Formación Particular Obligatoria	6. Departamento Sistemas de Información
7. Modalidad	8. Tipo de Asignatura: Curso-Taller	
9. Carga Horaria		
Teoría: 48 hrs	Práctica: 0 hrs	Total: 48 hrs	Créditos: 7
10. Trayectoria de la asignatura		

Contenido del Programa

11. Presentación
12.- Objetivos del programa Contar con los conocimientos necesarios en software y hardware para el procesamiento de grandes bases de datos en el contexto de economía-finanzas, negocios-mercadotecnia y políticas públicas
Objetivo General
13.-Contenido
Contenido temático 1. Introducción a grandes bases de datos y a la computación en la nube Fuentes de información 2. Fiabilidad de los datos 3. Tipologías y arquitecturas de un sistema de grandes bases de datos 4. Procesamiento por lotes (*Batch*) 5. Procesamiento en flujo (*Streaming*) 6. Procesamiento de grafos 7. Introducción a almacenamiento de los datos 8. Consistencia de los datos

Contenido desarrollado
Unidad 1. Introducción a grandes bases de datos y a la computación en la nube Fuentes de información
Objetivo particular de la unidad:
Desarrollo 1.1. Grandes bases de datos 1.1.1. Definición 1.1.2. Tipos de datos 1.1.3. Características 1.1.4. Las Vs de las grandes bases de datos 1.1.5. Aprovechamiento de grandes bases de datos 1.1.6. Generación de grandes bases de datos 1.1.7. Modelo de generación de grandes bases de datos 1.1.8. Manejo de grandes bases de datos 1.1.9. Valor del análisis de grandes bases de datos 1.1.10. Desafíos, Panorama y Tecnologías 1.2. Computación en la nube 1.2.1. Definición 1.2.2. Características esenciales 1.2.3. Modelos de servicios 1.2.3.1. Infraestructura como un servicio 1.2.3.2. Plataforma como un servicio 1.2.3.3. Software como un servicio

1.2.4. Modelos de implementación en la nube

1.2.4.1. Nube publica

1.2.4.2. Nube privada

1.2.4.3. Nube comunitaria

1.2.4.4. Nube híbrida

1.3. Soluciones industriales

1.3.1. Hadoop

1.3.2. Plataforma en la nuble de Google

1.3.3. Servicios de la Web de Amazon

1.3.4. Consola de administración de AWS

1.3.5. Microsoft Azure

Unidad 2 Tipologías y arquitecturas de un sistema de grandes bases de datos

Objetivos particulares de la unidad:

Conocer los Fundamentos tecnológicos de grandes bases de datos.

Identificar Arquitectura de un sistema de grandes bases de datos.

Aprender métodos de procesamiento distribuido.

Desarrollo

2.1 Fundamentos tecnológicos

2.1.1 Fundamentos tecnológicos

2.1.2 ¿Como procesamos toda esa información?

2.1.3 Computación científica

2.2 Arquitectura de un sistema de grandes bases de datos

2.2.1 Estructura general de un sistema de grandes bases de datos

2.2.2 Sistema de archivos

2.2.3 Sistema de cálculo distribuido

2.2.4 Gestor de recursos

2.2.5 *Stacks* de *software* para sistemas de grandes bases de datos

2.3 Escenarios de procesamiento distribuido

2.3.1 Procesamiento en *batch*

2.3.2 Procesamiento en *stream*

2.3.3 Procesamiento en grafos

2.3.4 Procesamiento en GPU

Unidad 3 Procesamiento por lotes (*Batch*)

Objetivos particulares de la unidad:

Conocer métodos de captura y preprocesamiento por lotes.

Aprender métodos de almacenamiento de datos estructurados.

Analizar datos estáticos.

Desarrollo

3.1 Captura y preprocesamiento por lotes

3.1.1 Conceptos básicos

3.1.2 Captura de datos estáticos

3.2 Almacenamiento de datos estructurados

3.2.1 Almacenamiento de datos masivos

3.2.2 Sistemas de ficheros distribuidos

3.2.3 Bases de datos *NoSQL*

3.3 Análisis de datos estáticos

3.3.1 *Apache Hadoop* y *MapReduce*

3.3.2 *Apache Spark*

Unidad 4 Procesamiento en flujo (*Streaming*)

Objetivo particular de la unidad:

Conocer métodos de captura y preprocesamiento de datos dinámicos.

Aprender métodos de almacenamiento de datos dinámicos.

Analizar datos dinámicos.

Desarrollo

4.1 Captura y preprocesamiento de datos dinámicos
- 4.1.1 Conceptos básicos
- 4.1.2 Captura de datos en *streaming*
- 4.1.3 Arquitectura de datos en *streaming*

4.2 Almacenamiento de datos dinámicos
- 4.2.1 Almacenamiento de datos dinámicos
- 4.2.2 Bases de datos en memoria

4.3 Análisis de datos dinámicos
- 4.3.1 Soluciones basadas en datos y basadas en tareas
- 4.3.2 Calculo *online* de valores estadísticos
- 4.3.3 Técnicas de resumen para el procesado aproximado de datos en flujo

Unidad 5 Procesamiento de grafos

Objetivos particulares de la unidad:

Conocer métodos de Representación y captura de grafos.

Aprender métodos de almacenamiento de grafos.

Analizar datos mediante técnicas de grafos.

Desarrollo:

5.1 Representación y captura de grafos

5.1.1 Conceptos básicos de grafos

5.1.2 Tipos de grafos

5.1.3 Captura de datos en formato de grafos

5.2 Almacenamiento de grafos

5.2.1 Almacenamiento en ficheros

5.2.2 Bases de datos *NoSQL* en grafo

6. Análisis de grafos

6.1 Procesamiento de grafos

6.2 Visualización de grafos

6.3 Herramientas para datos masivos

Unidad 6. Introducción a almacenamiento de los datos

6.1. Jerarquía de memoria

6.1.1. ROM

6.1.2. RAM

6.1.3. Almacenamiento secundario/respaldo

6.1.4. Discos duros

6.1.4.1. Grabación longitudinal

6.1.4.2. Almacenamiento de bits

6.1.4.3. Geometría y estructura

6.1.4.4. Acceso

6.1.5. Discos de estado sólido 1 de oct

6.1.6. Comparación HDD contra SSD

6.2. Matriz redundante de discos independientes (RAID)

6.2.1. Componentes de la matriz

6.2.2. Técnicas RAID

6.2.3. Implementaciones

7. Fiabilidad de los datos

7.1. Introducción a la replicación 8 de octubre

7.1.1. Problema de confiabilidad de los datos

7.1.2. Definición de replicación

7.1.2.1. Consistencia del sistema de archivos

7.1.2.2. Consistencia de la base de datos

7.1.3. Clasificación (local y remota)

7.1.4. Copia en primer acceso

7.1.5. Seguimiento de cambios en origen y destino

7.2. Introducción a los códigos de borrado (EC)

7.2.1. Bases de codificación de borrado

7.2.2. Códigos horizontales y verticales

7.2.3. Expresando Código con Generador Matriz

7.2.3.1. Codificación: Linux RAID-6

7.2.3.2. Aceleración de la codificación

7.2.3.3. Codificación: RDP

7.2.4. Aritmética para códigos de borrado

7.2.5. Decodificación con matrices generadoras

7.2.5.1. Códigos: Reed Solomon, EVENODD 1995, Código X 199,H y HDP

7.3. Replicación y EC en la nube 22 de octubre

7.3.1. Tres dimensiones en el almacenamiento en la nube

7.3.2. Replicación versus Codificación de borrado (RS)

7.3.3. Compensación fundamental

7.3.4. Códigos piramidales

7.3.4.1. Google GFS II y Microsoft Azure

7.3.5. Problema de recuperación en la nube

7.3.5.1. Optimizando la recuperación

7.3.5.2. Códigos Regeneradores

7.3.6. Combinación de dos códigos de borrado

8. Consistencia de los datos

8.1. Consistencia de datos y teorema CAP

8.1.1. Los sistemas de intercambio de datos de hoy

8.1.2. Propiedades fundamentales

8.1.2.1. Consistencia, disponibilidad, tolerancia a las particiones de red

8.1.3. Teorema CAP (Consistency, Availability and Partition Tolerance)

8.1.3.1. Definición

8.1.3.2. Perder particiones, perder disponibilidad, perder consistencia.

8.1.3.3. CAP en el Sistema de base de datos

8.1.4. Otro CAP: ACID (Atomicity, Consistency, Integrity, Durability)

8.1.5. Comparativa CAP y ACID

8.2. Protocolo de consenso: 2PC y 3PC

8.2.1. 2PC (Two Phase Commit Protocol): definición y problemática

8.2.2. 3PC (Three Phase Commit Protocol)

8.2.2.1. Definición

8.2.2.2. Solución de bloqueo

8.2.2.3. Especificaciones de manejo de tiempo de espera

8.2.2.4. Gestión de particiones y con particiones de red

8.2.2.5. Seguridad contra vivacidad.

8.3. Paxos

8.3.1. Definición, propiedades y desafíos

8.3.2. Mecanismos diferenciadores centrales e implementación

8.4. Chubby y Zookeeper

8.4.1. Chubby

8.4.1.1. Definición, estructura del sistema, lista de control de acceso,

cerraduras y secuencias, eventos, Apis, almacenamiento en caché

8.4.2. Zookeeper

8.4.2.1. Definición, ecosistema *Hadoop*, servicio, cliente API, garantías,

implementación, ambiente, ejemplo, aplicación

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 en la nube *(Cloud computing)*.

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

1. BIG DATA ARCHITEC´S HANDBOOK. Fahad, F. 2018 USA.
2. DATA CLOUD COMPUTING, DATA SCIENCE AND ENGINEERING. Lee, R. Big 2018 USA.
3. 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

Otros materiales

Cloud: herramientas para trabajar en la nube Autores: Celaya Luna, Ainoa ISBN: 9788490213858, 9781512949605 Editorial: Editorial ICB Año de Edición: 2014.

Computación en la nube para automatizar unidades de información. Revista Bibliotecas. Vol. 30, No. 1, 2012. Autores: Fernández Morales, Mynor ISBN: 1659328630105 Editorial: Red Universidad Nacional de Costa Rica. Año de Edición: 2012.

Computación en la Nube: estudio de herramientas orientadas a la industria 4.0. Lámpsakos, (20), 68-75 Autores: Fernández Ledesma, Javier Darío - Cadavid Nieto, Santiago - Ortiz Clavijo, Luis Felipe ISSN: 21454080 Editorial: Universidad Católica Luis Amigó Año de Edición: 2018.

Arquitectura para el aprovisionamiento dinámico de recursos computacionales. Autores: Vazquez Blanco, Constantino - Ignacio Martín LLorente - Eduardo Huedo Cuesta ISBN: 71607131710 Editorial: Universidad Complutense de Madrid Año de Edición: 2012

18.- Perfil del profesor

19.- Nombre de los profesores que imparten la materia

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

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