Datos Masivos I

Syllabus

Febrero 2023						
Do.	**Lu.**	**Ma.**	**Mi.**	**Ju.**	**Vi.**	**Sá.**
			1 COV, Resp. Curso. Environment install PySpark package	**2** Introducción (P)	**3** Introducción (P)	**4**
5	**6**	**7**	**8** Almacenamiento Distribuido (P)	**9** Map – Reduce (P)	**10** PySpark (T)	**11**
12	**13**	**14**	**15** Modelo Programac. Map – Reduce (P)	**16** Algoritmos Map – Reduce (P) **TR 1**	**17** Map – Reduce (T) Lazy Evaluation(T)	**18**
19	**20**	**21**	**22** Extensiones Map – Reduce (P)	**23** Teoría de la Complejidad (P)	**24** Operaciones Básicas (T)	**25**
26	**27**	**28**				

P = Presentación **T = Taller / Práctica** **TR = Tarea** **PF = Proyecto Final**

Marzo 2023

Do.	Lu.	Ma.	Mi.	Ju.	Vi.	Sá.
			1 Modelo Costo – Comunicación (P) **Entrega TR 1**	**2** Modelo Costo – Comunicación (P)	**3** Medidas Similitud y Distancia (P)	**4**
5	**6**	**7**	**8** Medidas Similitud y Distancia (P)	**9** Resúmenes de Conjuntos (P)	**10** Función Hash (T) Índice inverso (T)	**11**
12	**13**	**14**	**15** Resúmenes de Conjuntos (P)	**16** Funciones Hash (P) **TR 2**	**17** Min Hash (T) Buscador (T)	**18**
19	**20**	**21**	**22** Funciones Hash (P)	**23** Funciones Hash (P)	**24** MinHash PySpk(T) MinHash PySpk(T) MinHash PySpk(T)	**25**
26	**27**	**28**	**29** Flujos de Datos (P) **Entrega TR 2**	**30** Flujos de Datos (P)	**31** Examen (E)	

P = Presentación T = Taller / Práctica TR = Tarea PF = Proyecto Final

Abril 2023

Do.	Lu.	Ma.	Mi.	Ju.	Vi.	Sá.
						1
2	3	4	5 Muestreo – Filtrado (P)	6 Muestreo – Filtrado (P)	7 Muestreo (T) Streaming (T)	8
9	10	11	12 Conteo (P)	13 Conteo (P) **TR 3**	14 Filtro Bloom (T) Elementos Dist. (T)	15
16	17	18	19 Momentos (P)	20 Momentos (P)	21 Momentos (T) Explicar PF (Kaggle, etc.)	22
23	24	25	26 Memoria Externa (P) **Entrega TR 3**	27 Memoria Externa (P)	28 Kafka 1 (T) Kafka 2 (T)	29
30						

P = Presentación T = Taller / Práctica TR = Tarea PF = Proyecto Final

Mayo 2023

Do.	Lu.	Ma.	Mi.	Ju.	Vi.	Sá.
	1	**2**	**3** Memoria Externa (P)	**4** Presentación de Proyecto (bases, plan, datasets, etc.)	**5** Presentación de PF (modelo, plan, datasets, etc.)	**6**
7	**8**	**9**	**10** Modelo Inconscient de Caché (P)	**11** Modelo Inconscient de Caché (P)	**12** Kafka 3 (T) Kafka 4 (T) Kafka 5 (T)	**13**
14	**15**	**16**	**17** Ordenamiento y Búsqueda (P)	**18** Ordenamiento y Búsqueda (P)	**19** PF: Dudas	**20**
21	**22**	**23**	**24** Entrega PF	**25** Entrega PF	**26** Entrega PF	**27**
28	**29** Exámenes Finales I	**30** Exámenes Finales I	**31** Exámenes Finales I			

P = Presentación **T = Taller / Práctica** **TR = Tarea** **PF = Proyecto Final**

Junio 2023

Do.	Lu.	Ma.	Mi.	Ju.	Vi.	Sá.
				1 Exámenes Finales I	2 Exámenes Finales I	3
4	5 Exámenes Finales II	6 Exámenes Finales II	7 Exámenes Finales II	8 Exámenes Finales II	9 Exámenes Finales II	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

P = Presentación T = Taller / Práctica TR = Tarea PF = Proyecto Final

Contenido Datos Masivos I

Objetivos del Curso.

Formular las estructuras de datos y algoritmos para el almacenamiento, manejo, organización y análisis de datos masivos.

Objetivos Específicos.

Producir algoritmos escalables para búsqueda de elementos similares en datos masivos.

Usar algoritmos para procesar y analizar flujos de datos.

Generar algoritmos y estructuras de datos eficientes que operan directamente en datos almacenados en memoria externa.

Diseñar algoritmos que usen de forma eficiente el modelo de programación de mapeo y reducción para procesar datos masivos.

1. Conceptos básicos

1.1 Definición y características

1.2 Generación, procedencia y preparación de datos

1.3 Consideraciones estadísticas y computacionales de los datos masivos

1.4 El principio de Bonferroni

1.5 Privacidad y riesgo

1.6 Modelos de computación para datos masivos

2. Modelo de mapeo y reducción

2.1 Sistema de almacenamiento y procesamiento distribuido

2.2 Modelo de programación

2.3 Algoritmos con el modelo de mapeo y reducción

2.4 Extensiones

2.5 El modelo costo-comunicación

2.6 Teoría de la complejidad para el modelo de mapeo y reducción

3. Búsqueda de elementos similares

3.1 Medidas de similitud

3.2 Resúmenes de conjuntos con preservación de similitud

3.3 Funciones hash sensibles a la localidad

3.4 Métodos para altos grados de similitud

3.5 Aplicaciones

4. Algoritmos para flujos de datos

4.1 El modelo de datos en flujo

4.2 Muestreo

4.3 Filtrado

4.4 Conteo

4.5 Estimación de momentos

4.6 Ventanas deslizantes

5. Algoritmos de memoria externa

5.1 Modelo de memoria externa

5.2 Modelo de caché inconsciente

5.2 Cotas fundamentales de operaciones de entrada y salida

5.3 Escaneo

5.4 Ordenamiento

5.5 Búsqueda

5.6 Estructuras de datos estáticos y dinámicos

Evaluación

Proyecto (45 %)

Tareas (30 %)

Exámenes (15 %)

Participación (10 %) (Asistencia)

Envío de tareas al correo (raul.galindo.hernandez@comunidad.unam.mx) hasta el último segundo del día.

Poner en Asunto: Tarea # NombreCompleto

Ejemplo: Tarea 1 Rodrigo Velez Pérez

Proyecto: Presentaciones individuales

1. Propuesta
2. Proyecto Final (Inglés)

Examen: Presencial y en papel.

Se requiere entregar proyecto para tener derecho al primer final. Si no se entrega proyecto, se presentará solo el segundo final.

Ambiente de programación.

- Instalación de Anaconda Navigator
- Instalación de paquetes (PySpark: pip install pyspark, conda install pyspark)

Bibliografía del curso

- Jure Leskovec, Anand Rajaraman and Jeffrey D. Ullman. Mining of Massive Datasets. Second Edition. Cambridge University Press, 2014.
- Charu C. Aggarwal. Data Mining. Springer International Publishing, 2015.
- Jeffrey Vitter. Algorithms and Data Structures for External Memory. Now Foundations and Trends, 2008.

Consideraciones Generales

- Asistencia será tomada en cuenta como máximo 10 min después de comenzar la clase.
- Dudas y preguntas se deben realizar durante la clase, no después.
- Respeto en la clase.
- Cubrebocas (if possible). Si llegan a sentir algún síntoma(s) favor de mandarme correo a la brevedad y revisar que se puede hacer en su caso (por supuesto no afectará su evaluación).
- Respuestas a sus correos pueden ser rápidas y cortas.
- Búsquedas en inglés.
- Correcto nombrado de las variables de sus tareas y proyecto.
- NO Chat GPT