COM 12103: FUENTES DE DATOS

Información del profesor

Nombre: Luis Manuel Román García
Correo: luis.roman@itam.mx

Información de la clase

Fechas: 24-01-2022 – 25-05-2022
Horario: MA-JU. 17:30 – 19:00
Salón: RHCC102

Descripción del curso

Todo proceso científico que tenga como finalidad describir o modelar un fenómeno observable, contrastar una hipótesis o simplemente explorar información en búsqueda de patrones interesantes, requiere el uso de técnicas eficientes para extraer, manipular, almacenar y consumir datos. En este curso abordaremos distintas tecnologías involucradas en este proceso, las exploraremos de manera individual así como embebidas dentro de un encuadre general de análisis de datos. Dada la flexibilidad que confiere conceptualizar toda fuente de datos como unidades informacionales almacenadas en texto, se hará un gran énfasis en técnicas que nos permitan el procesamiento eficiente de este tipo de formatos. Finalmente, y dado que la manipulación de datos masivos se ha vuelto la norma más que la excepción, buscaremos que las metodologías exploradas en clase sean soluciones integradas que escalen de manera horizontal con el volumen de datos.

Objetivos del curso

Después de este curso, debes ser capaz de...

- Usar la terminal y lenguajes de 'scripting' como Sed y AWK para generar flujos de extracción, procesamiento y limpieza de información.
- Explorar y analizar fuentes más sofisticadas de información con Python y R.
- Explotar información en distintos formatos: csv, json, web; y múltiples tipos de almacenamiento: SQL, NoSQL.
- Escalar procesos a escenarios de datos masivos con Pyspark.
- Generar soluciones empaquetadas con mejores prácticas de mantenimiento y documentación de software.

Libros de texto, & Software

Robbins, Arnold, and Nelson HF Beebe. Classic Shell Scripting: Hidden Commands that Unlock the Power of Unix. " O'Reilly Media, Inc.", 2005.

Dougherty, Dale, and Arnold Robbins. sed & awk: UNIX Power Tools. " O'Reilly Media, Inc.", 1997.

Wickham, Hadley, and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.", 2016.

Wickham, Hadley. Advanced r. CRC press, 2019.

Harrison, Matt, and Theodore Petrou. Pandas 1. x Cookbook: Practical recipes for scientific computing, time series analysis, and exploratory data analysis using Python. Packt Publishing Ltd, 2020.

Asistencia y participación

Es escencial para el éxito en el curso la asistencia y participación en las discusiones de clase. La participación es un componente importante de la evaluación.

Tareas & proyectos

Este curso busca maximizar la aplicabilidad de los conceptos vistos en clase, por tanto, cada módulo irá acompañado de un proyecto que se realizará en equipos. Las tareas resultaran de ejericios que se dejen en clase y su realización es opcional aunque fuertemente sugerida.

Examen final

El examen final comprenderá un caso de uso que cubra el total de los conceptos vistos en clase.

Evaluación

La calificación final se compone de la siguiente forma:

Participación	10 pts
Proyecto 1	15 pts
Proyecto 2	20 pts
Proyecto 3	25 pts
Proyecto 4	30 pts
Examen final	25 pts

Notar que el curso comprende de 125 puntos, por tanto, es necesario acumular al menos 75 puntos para aprobar. Los equipos que hayan realizado todas las tareas y cuenten con el total de asistencias durante el curso tendrán un punto extra sobre la calificación final.

Calendario curso

El siguiente es un calendario *tentativo* para el curso.

Semana	Fecha	Proyecto	Fecha entrega
1 - 3	24/01 - 07/02	Proyecto 1	13/02 11:59 pm (CST)
4 - 7	14/02 - 28/02	Proyecto 2	06/03 11:59 pm (CST)
7 - 10	03/07 - 03/21	Proyecto 3	27/03 11:59 pm (CST)
10 - 13	28/03 - 18/04	Proyecto 4	24/04 11:59 pm (CST)