

COM 12103: Fuentes de Datos

Información del Profesor

Nombre: Mario Vazquez Corte

Correo: mario.vazquez.corte@itam.mx

Correo alternativo: vazcorm@gmail.com

Informacion de la Clase

Fecha de Inicio: 9 de Enero 2023

Fecha Final: 10 de Mayo 2023

Horario: Lunes y Miercoles de 17:30 a 19:00

Salon:

Repositorio del Grupo

https://github.com/sonder-art/fdd_prim_2023



Descripción del Curso

Todas las hipótesis científicas necesitan datos para ser evaluadas y estudiadas. Como científicos de datos es importante que aprendamos a extraer, almacenar, manipular, utilizar,

visualizar y consumir correctamente los datos. Esto incluye la recopilación o generación de datos y su uso final por parte de otras personas.

Durante el curso, contaremos con la visita de expertxs de la industria y la academia que nos contarán sobre su día a día y nos enseñarán algunas de las herramientas que han desarrollado.

Objetivos del Curso

Nos centraremos en comprender las particularidades y generalidades de la manipulación técnica, teorica y aplicada de datos, así como las mejores prácticas y herramientas para hacerlo. Tener una arquitectura de datos correctamente mapeada y con las mejores prácticas es el primer paso para cualquier proyecto de datos que queramos llevar a cabo.

1. Conocer y utilizar correctamente Ubuntu (sistemas operativos basados en Linux) usando herramientas open source en la medida de lo posible.
 - a. Uso de Terminal, comandos basicos, scripts y shell/bash.
 - b. Ssh
 - c. VScode
 - d. Personalizacion de nuestro entorno de desarrollo
2. Manipulacion de datos en Terminal
 - a. Grep
 - b. AWK
3. Buenas practicas de programacion
 - a. Pyenv, pip y conda
 - b. Docker
 - c. Pytest y Typing
4. Manejo de Librerías de Datos en Python
 - a. Numpy
 - b. Pandas
 - c. Pyspark
 - d. Dask
5. Manipulación de Datos
 - a. SQL
 - b. MongoDB
 - c. Csv, json y web

Libros de Texto y Software

Newham, C. (2005). *Learning the bash shell: Unix shell programming*. " O'Reilly Media, Inc.", 3d Edition.

Rioux, Jonathan. *Data Analysis with Python and Pyspark*. Manning Publications, 2022.

McKinney, W. (2017). *Pandas in Action*. Shelter Island, NY: Manning.

Wickham, H. (2014). Tidy data. New York, NY: Springer.

Reis, J., & Housley, M. (n.d.). Fundamentals of data engineering. Sebastopol, CA: O'Reilly Media, Inc.

Tareas y Proyectos

Durante el curso realizaremos varias tareas y algunos proyectos.

Proyecto de Instalacion: Crear un Raspberry-Pi con dual boot

Mas proyectos y tareas por definir

Evaluacion

Tareas	30 pts
Proyectos	50 pts
Proyecto/Examen final	20 pts
Participacion	10 pts