Universidad Tecnológica de la Mixteca

Clave DGP: 200089

Doctorado en Modelación Matemática 00059

PROGRAMA DE ESTUDIOS

NOMBRE DE LA ASIGNATURA
Herramientas computacionales para ciencia de datos

SEMESTRE	CLAVE DE LA ASIGNATURA	TOTAL DE HORAS
Optativa	**292917**	**80**

OBJETIVO(S) GENERAL(ES) DE LA ASIGNATURA
Que el estudiante domine el manejo de las herramientas computacionales básicas para el manejo de grandes cantidades de datos y el control de versiones de programas.

TEMAS Y SUBTEMAS

1. Manipulación de datos desde la línea de comandos
- 1.1 Archivos y directorios
- 1.2 Filosofía UNIX
- 1.3 Movimiento en la terminal
- 1.4 Pipes y Redirecciones
- 1.5 Comandos útiles: seq, tr, wc, head, tail, cat, uniq, cut, sort
- 1.6 Jobs
- 1.7 file, iconv, od
- 1.8 Expresiones regulares
- 1.9 Analizando datos: grep, awk, sed
- 1.10 Bash programming

2. Controladores de versiones
- 2.1 Configuración de git
- 2.2 Solo Workflow
- 2.3 Github Workflow
- 2.4 Branches
- 2.5 Push, Pull y Pull request
- 2.6 Merges
- 2.7 Tags

3. Desarrollo de software para productos de datos
- 3.1 Buenas prácticas en programación
- 3.2 Paradigmas de programación: Procedural, Orientado a Objetos, Funcional, Lógico/Declarativo, Spaghetti code
- 3.3 Diseño semántico
- 3.5 Pruebas en Ciencia de datos
- 3.6 Reproducibilidad: Buenas prácticas
- 3.7 Reproducibilidad: Ambientes
- 3.8 Reproducibilidad: Máquinas virtuales y Contenedores

4. Almacenar datos: Bases de datos y SQL
- 4.1 ¿Por qué usarlas?
- 4.2 Bases de datos relacionales
- 4.3 Data model
- 4.4 Introducción a SQL
- 4.5 SQL Avanzado
- 4.6 SQL para Ciencia de Datos: Feature engineering
- 4.7 Diseño de esquemas para un producto de datos
- 4.8 SQL para datos espaciales: GIS
- 4.9 Comparación con bases NOSQL

5. Escalar ejecución de código
- 5.1 Limitaciones
- 5.2 Tipos de escalamiento
- 5.3 GNU/Parallel usando CPUs

5.4 Solo Python: Dask
5.5 Usando la herramienta adecuada: diseño de solución
5.6 Pipelines en productos de datos

6. Escalar en la nube
7.1 La nube: conceptos y herramientas
7.2 GNU/Parallel usando la nube
7.3 Ejecución batch
7.4 Ejecución on-line
7.5 Serverless

ACTIVIDADES DE APRENDIZAJE

Sesiones dirigidas por parte del profesor, con prácticas de los programas especializados en cada unidad. Se recomienda que la unidad 4 sea abordada con MySQL o, en su defecto, con PostgreSQL. Así también que la unidad 6 sea analizada con Microsoft Azure o con Amazon Web Services (AWS). Se recomienda ampliamente impartir el curso en un laboratorio con equipo de cómputo disponible para cada estudiante.

CRITERIOS Y PROCEDIMIENTOS DE EVALUACIÓN Y ACREDITACIÓN

Se aplican por lo menos tres exámenes parciales cuyo promedio equivale al 50% de la calificación final, el 50% restante se obtiene de un examen final. Otras actividades que se consideran para la evaluación son las participaciones en clase, asistencias a clases y el cumplimiento de tareas.

BIBLIOGRAFÍA

Básica:

1. J. Bird, A basic UNIX tutorial, Thee Idaho State University Computer Center.
2. R. Somasundaram, Git: Version Control for Everyone Beginner's Guide, Pack Publishing.
3. A. Beaulieu, Learning SQL, O'Reilly.
4. J. C. Daniel, Data Science With Python And Dask, Manning Publications, 2019.

Consulta:

1. Henry Li, Introduction to Windows Azure: an introduction to cloud computing using Microsoft Windows Azure, Series: The expert's voice in web development, Apress; Distributed by Springer-Verlag New York, 2009.

PERFIL PROFESIONAL DEL DOCENTE

Estudios de Doctorado en Matemáticas o en Estadística.

Vo.Bo
DR. JOSÉ ANIBAL ARIAS AGUILAR
JEFE DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO

DIVISION DE ESTUDIOS DE POSGRADO

AUTORIZÓ
DR. RAFAEL MARTÍNEZ MARTÍNEZ
VICE-RECTOR ACADÉMICO

VICE-RECTORIA ACADÉMICA