MINE 4213- Sistemas Intensivos en Datos

Programa del Curso

Información General

Profesores: Christian Ariza, Yachay Tolosa
Infraestructura práctica: Apache Spark, MongoDB Atlas, Apache Kafka, Unity Catalog, MLFlow, entre otros.

Descripción del curso

Este curso se enfoca en la arquitectura de datos y las habilidades de ingeniería de datos necesarias para el procesamiento distribuido a gran escala. Los estudiantes utilizarán herramientas del ecosistema de Apache Spark para construir, optimizar y gestionar pipelines de datos. A lo largo del curso, se trabajará con conjuntos de datos representativos de escenarios reales y se aplicarán mejores prácticas de la industria, con énfasis en gobernanza, formatos de almacenamiento y generación de características (feature engineering).

Al finalizar el curso, los estudiantes serán capaces de:

1. Procesar datos a gran escala utilizando Apache Spark.
2. Diseñar arquitecturas de datos para aplicaciones en entornos productivos.
3. Implementar y optimizar pipelines de datos con Delta Lake, gestionar su gobernanza y control de acceso con Unity Catalog, y administrar la experimentación y el ciclo de vida de modelos de machine learning con MLflow.
4. Gestionar la ingeniería de características y la trazabilidad de los datos para entrenar y desplegar modelos de machine learning de manera reproducible.

Estructura del curso

Módulo 1: Introducción a Soluciones Intensivas en Datos (2 semanas)

- Introducción al curso y configuración de Databricks
- Fundamentos de computación distribuida y procesamiento en paralelo
- Visión general de arquitecturas de datos (procesamiento por lotes vs. en tiempo real)
- Repaso de modelado relacional y no relacional

Módulo 2: Ingeniería de Datos con Spark (3 semanas)

- Arquitectura de Spark y modelo de ejecución
- Ingesta y transformación de datos con Spark

- Particionamiento y optimización de rendimiento
- Práctica: Carga y transformación de grandes volúmenes de datos

Modulo 3: Procesamiento de datos en tiempo real (2 semanas)

- Arquitecturas de para Streaming de datos (Fast Data)
- Kafka
- Spark Structured Streaming (y otros motores de procesamiento de datos en streaming)

Examen 1

Módulo 4: Formatos de Almacenamiento y Gobernanza de Datos (3 semanas)

- Delta Lake: Transacciones ACID y time travel
- Parquet vs. Avro vs. ORC: Selección del formato adecuado
- Introducción a **Unity Catalog** para gobernanza y seguridad
- Práctica: Implementación de políticas de gobernanza en Unity Catalog

Módulo 5: Orquestación y automatización (2 semanas)

- Orquestación de workflows con **Databricks Workflows**
- Automatización de procesos ETL
- Mejores prácticas de monitoreo y depuración
- Práctica: Implementación de un pipeline ETL de extremo a extremo

Módulo 6: Gestión del Ciclo de Vida de Modelos de Machine Learning (2 semanas)

- Introducción a **MLFlow** (seguimiento, modelos, registro)
- Feature engineering y feature stores
- Gestión del ciclo de vida de modelos en producción
- Práctica: Uso de MLFlow para el seguimiento de experimentos

Examen Final

Proyecto Final (Transversal a Todo el Curso)

- Equipos de **3 estudiantes**, con dos entregas de proyecto (parcial y final)

Evaluación y Calificación

- **Exámenes (60%)**: Dos exámenes individuales sobre teoría y práctica
- **Proyecto Final (30%)**: Trabajo en equipo resolviendo un problema concreto

- **Cuestionarios y Trabajos en Clase (10%)**: Asignaciones en clase y participación

---

Bibliografía Recomendada

1. *Designing Data-Intensive Applications* (2da edición en early release)
2. *Kafka: The Definitive Guide* (2da edición)
3. *Fast Data Architectures for Streaming Applications*
4. *Learning Spark* (2da edición)
5. Delta Lake: The Definitive Guide
6. Documentación oficial de Databricks, Airflow y Argo Workflows