# Gestión de Datos Masivos

## Información del Curso

Profesor: Alejandro Vasquez

Correo: [Alejandro.vasquez@pascualbravo.edu.co](mailto:Alejandro.vasquez@pascualbravo.edu.co)

Horario atención: L-W 6-8 pm.

Periodo: 2025-II

## Textos de ayuda:

1. **“Fundamentals of Data Engineering” (2022)** – Joe Reis, Matt Housley.
2. **“Learning Spark: Lightning-Fast Data Analytics” (2nd Ed., 2020)** – Jules Damji, Brooke Wenig, Tathagata Das, Denny Lee.
3. **“Data Analytics with Hadoop: An Introduction for Data Scientists” (2nd Ed., 2021)** – Benjamin Bengfort, Jenny Kim, Michelle Brush.
4. **“Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing” (Updated Ed., 2021)** – Tyler Akidau, Slava Chernyak, Reuven Lax.

## Tipo de Instrucción

Clases magistrales, discusión guiada, resolución de problemas, laboratorios de programación, estudio de casos, desarrollo de un proyecto aplicado.

## Cronograma Tentativo

**Parte 1 – Fundamentos y ciclo de vida (Semanas 1–5)**

| **Semana** | **Tema** |
| --- | --- |
| 1 | Introducción a Big Data, características (5V), casos de uso |
| 2 | Datos distribuidos y escalabilidad |
| 3 | Costos computacionales de almacenamiento y consulta |
| 4 | Ciclo de vida de datos (adquisición, almacenamiento, análisis, archivado) |
| 5 | Herramientas y entornos de Big Data |

**Parte 2 – Ecosistemas y arquitecturas (Semanas 6–10)**

| **Semana** | **Tema** |
| --- | --- |
| 6 | Tipos de fuentes de datos: estructurados, no estructurados, semi-estructurados |
| 7 | Ingesta y preparación de datos masivos |
| 8 | Almacenamiento y flujos de trabajo |
| 9 | Arquitecturas de referencia (simple, Lambda, Kappa) |
| 10 | Ecosistema Hadoop: HDFS, YARN, Hive, Pig |

**Parte 3 – Procesamiento y arquitecturas avanzadas (Semanas 11–16)**

| **Semana** | **Tema** |
| --- | --- |
| 11 | MapReduce: fundamentos y ejecución |
| 12 | Sistemas de datos distribuidos Hadoop |
| 13 | Arquitectura Lambda |
| 14 | Arquitectura Kappa |
| 15 | Almacenamiento y flujos batch y streaming |
| 16 | Presentación final de proyectos |

## Evaluación – Proyecto de Aula

* **Entrega 1 (30%, Semana 6)**  
  Diseño del caso de estudio y plan de arquitectura de datos (fuentes, ingesta, almacenamiento, procesamiento).
* **Entrega 2 (30%, Semana 11)**  
  Implementación parcial: ingesta, almacenamiento y flujo batch con Spark.
* **Entrega 3 (40%, Semana 16)**  
  Implementación final incorporando batch y/o streaming, documentación y defensa oral.