Semana 3 — plan de sesión (20 jun 2026)

Electiva Big Data · Maestría en Estadística Aplicada · UDENAR · A-2026

Sábado 20 jun 2026 · 07:00–13:00 (hora Colombia) · Lecciones **L5** (Ingesta y flujos) + **L6** (Analítica y visualización)

Antes del sábado

- Cuadernos individuales **L5** y **L6** ejecutados (Colab).
- Videos de introducción a los temas de la semana.
- Diario personal semana 3 (`learning-journal-week-3.docx`).

Cronograma del sábado

Hora (CO)	Duración	Actividad
07:00–07:15	15 min	Apertura
07:15–08:00	45 min	Discusión de lecturas — en su grupo
08:00–08:30	30 min	Discusión plenaria — lecturas
08:30–08:45	15 min	Descanso
08:45–10:45	2 h	L5 + L6 + cuaderno grupal en equipo
10:45–11:15	30 min	Plenaria — barreras de programación y datos
11:15–11:30	15 min	Descanso
11:30–12:45	1 h 15 min	Proyecto — capas de evidencia
12:45–13:00	15 min	Cierre

Discusión de lecturas (45 min grupo + 30 min plenaria)

En su grupo (45 min): discutan las preguntas del **diario** (Parte A y B) y las preguntas abajo.

Lectura	Pregunta para discutir
Zaharia et al. (2016)	¿Qué problema unifica Spark? Compare con lo que ya usó (MapReduce, DuckDB, Polars, lakehouse Parquet).
Jarrahi et al. (2023)	¿Qué critica el enfoque centrado en el modelo? ¿Qué principio de DCAI ya practica su grupo en el proyecto?

Plenaria — síntesis (C1–C3)

Pregunta	Síntesis
C1	Spark unifica batch/stream/ML, DCAI exige calidad del dato, el curso separa **capa oficial (GEIH)** y **contexto mediático**. ¿Cómo encajan en **un** tablero sin mezclar veracidades?
C2	¿Qué decisiones de su proyecto reflejan un enfoque centrado en los datos?
C3	Modelo **lineal vs red pequeña** *(teoría / diapositivas L6)*: ¿cuál mostraría a quien decide y cuál solo a un analista? *(En el cuaderno usamos solo regresión lineal.)*

Plenaria (30 min): un portavoz por grupo (máx. 5 min).

Cuaderno grupal L5 + L6 (2 h, en equipo)

Todo el grupo en un mismo Colab (`week-3-group`).

Ejercicio	Qué hace el grupo
1 (L5)	Capa **curated** GEIH 2022–2025 + `audit.jsonl` + `schema_contract.json` + flujo **Prefect** (ETL real)
2 (L6)	**Regresión lineal** con train/val/prueba + ≥2 gráficos + **tablero** sobre agregados oficiales *(el contraste lineal vs MLP es solo en diapositivas / video teórico)*

Nota: `audit.jsonl` y `schema_contract.json` en **Drive**; el ZIP lleva solo cuaderno + `manifest.json` (resumen de gobernanza). La ingesta de noticias/Kafka es **individual** (L5); no va en el ZIP grupal.

Entrega ZIP grupal **martes 23 jun 2026, 23:59** (detalle en el hub semana 3 y en Moodle).

Plenaria — barreras de programación y datos (30 min)

Discutimos posibles barreras encontradas en la implementación y cuadernos de esta semana.

Proyecto y analítica (1 h 15 min en grupo)

Avancen la **pregunta de decisión** y las **capas de evidencia** (oficial GEIH vs contexto) en el PDF de requerimientos.

Presentación final (sáb 27 jun): ~20 min (6 diapositivas + **demo en vivo**) + **~10 min** preguntas. Plantilla en el hub semana 3.

Entregas semana 3 (después del sábado)

Entrega	Plazo (Colombia)
ZIP grupal: `week-3-group-<group_id>.zip` (cuaderno + `manifest.json`; bitácora y contrato en Drive)	**Martes 23 jun 2026, 23:59**
Reflexión individual: `week-3-reflection-<student>.pdf`	**Miércoles 24 jun 2026, 23:59**