EXTRACCIÓN Y GESTIÓN DE DATOS MASIVOS 2025-2

PROYECTO GRUPAL

DESCRIPCIÓN

Los estudiantes deberán desarrollar un proyecto en grupo que involucre el uso de Apache Spark para procesar y analizar grandes volúmenes de datos. El proyecto debe abordar un problema real o simulado que involucre extracción, procesamiento y análisis de datos, demostrando la capacidad de Apache Spark para manejar grandes volúmenes de información de forma eficiente.

OBJETIVOS DE APRENDIZAJE

- Demostrar conocimiento y uso de herramientas de Apache Sparrk (Spark Core, Spark SQL, Spark Streaming, Spark MLlib, GraphX).
- Documentar y comunicar (exponer) los resultados de forma efectiva.

ETAPAS Y ENTREGABLES

Etapa 1

Título: Ideas de proyecto

Descripción: Cada grupo debe plantear tres ideas de proyecto. Por cada idea se debe definir contexto, motivación (problema a resolver), objetivos de análisis (información o conocimiento que se desea extraer), datos a emplear, y herramientas a emplear. El tamaño de los datos a emplear debe ser superior a 1GB (en formato CSV). Dentro de las herramientas empleadas, será obligatorio el uso de SparkMLlib o GraphX.

Entregables:

- Diapositivas (grupo1-etapa1.pptx). Exposición de 10 minutos.

Fecha de entrega: 3 y 5 de noviembre del 2025

Etapa 2

Título: Exploración de datos y herramientas

Descripción: Cada grupo debe realizar un análisis exploratorio de los datos, para demostrar que son apropiados para el proyecto, tanto en términos de calidad como cantidad. Además, el equipo debe realizar una demostración del funcionamiento de las herramientas que se usarán en el proyecto. Esto implica ejecutar un flujo de extracción, procesamiento y análisis, con una porción pequeña de datos. Las herramientas deben funcionar en un ambiente de cluster compuesto de tres nodos (1 master y 2 workers).

Entregables:

- Diapositivas (grupo1-etapa2.pptx): Mostrando información básica y estadísticas sobre los datos. Exposición de 5 minutos.
- Video (grupo1-etapa2.mp4): mostrando el flujo de extracción, procesamiento y análisis. Duración de 5 minutos.

Fecha de entrega: 10 y 12 de noviembre del 2025

Etapa 3

Título: Preparación del experimento

Descripción: Cada grupo debe analizar, diseñar, y ejecutar un experimento que implique la extracción, procesamiento y análisis de datos. El experimento debe estar asociado a extraer información o conocimiento relevante desde los datos elegidos por el grupo.
Entregables:

- Video (grupo1-etapa3.mp4): debe mostrar el flujo de extracción, procesamiento y análisis. Duración de 5 minutos.

Fecha de entrega: 17 y 19 de noviembre del 2025

Etapa 4
Título: Evaluación de pares
Descripción: Cada grupo debe elaborar un documento con instrucciones detalladas (paso a paso) de como ejecutar el experimento definido en la Etapa 3. Un segundo grupo ejecutará y evaluará el experimento.
Entregables:

- Documento con instrucciones (grupo1-etapa4.docx).

Fecha de entrega: 24 y 26 de noviembre del 2025

Etapa 5
Título: Evaluación de eficiencia
Descripción: Cada grupo debe evaluar la eficiencia de las herramientas en 3 configuraciones de cluster: (i) 2 workers; (ii) 4 workers; (iii) 8 workers. Esto implica ejecutar el workflow definido en la etapa 3, en cada una de las configuraciones indicadas. Se debe medir el tiempo de ejecución de cada una de las fases de procesamiento (extracción, procesamiento, análisis), por separado, y calcular el tiempo total.
Entregables:

- Video (grupo1-etapa5.mp4): debe mostrar la ejecución del workflow en cada configuración. Duración de 5 minutos.
- Diapositivas (grupo1-etapa5.pptx): debe incluir gráficos que permiten comparar los tiempos de ejecución de las distintas fases del workflow, además del tiempo global. Exposición de 5 minutos.

Fecha de entrega: 1 y 3 de diciembre del 2025

PONDERACIÓN

- Presentación Oral (30% de la nota del curso): Promedio de las presentaciones de las Etapas 1, 2, 3 y 5.
- Informe de proyecto (30% de la nota del curso): Documento de la Etapa 4.

REQUISITOS

- Cada grupo debe estar compuesto por 3 estudiantes (2 en caso justificado).
- Los integrantes de cada grupo deben auto-organizarse y distribuirse las tareas.
- Cuando un grupo deba exponer, el profesor será quien elija al expositor. Además, todos los integrantes del grupo deberán estar presentes en la exposición.
- Los experimentos deben funcionar con Spark 4.0.1.
- Los experimentos deben ejecutarse en máquinas virtuales del cluster empleado en el curso.