Orientación del Proyecto Final

Considerando que los equipos ya cuentan con la **infraestructura configurada**, la **carga inicial de datos** y la **arquitectura definida**, esta fase final se centrará en el procesamiento masivo, la optimización, la analítica avanzada y la visualización integral (tanto de resultados como de rendimiento).

Componentes del Proyecto Final

A. Procesamiento Masivo y Lógica de Negocio

- **Ejecución Completa:** Procesamiento del dataset completo (o un volumen significativo que justifique el uso de grandes datos).
- **Complejidad Algorítmica:** Implementación de la lógica central definida en los objetivos. Esto puede incluir:
  - Modelos de Machine Learning.
  - Consultas complejas con agregaciones y ventanas de tiempo.
  - Procesamiento de grafos o flujos de datos en tiempo real.

B. Visualización de Resultados

El sistema debe tener una capa de presentación que responda a la pregunta: *¿Qué conocimiento se extrajo de los datos?*

- **Dashboard:** Creación de un tablero (usando herramientas como PowerBI, Tableau, Grafana o Streamlit conectado a los datos procesados.
- **Interpretación:** Los gráficos deben ser pertinentes al objetivo del proyecto (ej. mapas de calor, líneas de tendencia, clasificación de entidades).

C. Visualización de Métricas del Clúster

Requisito indispensable. Se debe evidenciar el comportamiento de la infraestructura durante el procesamiento.

- **Monitoreo de Recursos:** Gráficos o capturas que muestren el consumo de CPU, RAM y uso de disco en los nodos del clúster (HDFS/YARN) durante la ejecución de los jobs.
- **Análisis de Tiempos:** Comparativa de tiempos de ejecución (ej. antes vs. después de alguna optimización, o comparación escalonada según volumen de datos).
- **Identificación de Cuellos de Botella:** Explicar si hubo saturación de red, falta de memoria en los ejecutores, etc.

Entregables

El proyecto final requiere de tres elementos obligatorios para su evaluación:

1. Informe Técnico Final

Documento que integre lo realizado en el semi-proyecto con los resultados finales. Debe contener:

- **Arquitectura Final:** Diagrama actualizado del pipeline si hubo cambios durante la implementación.
- **Metodología de Procesamiento:** Explicación técnica de las transformaciones y algoritmos aplicados (tampoco tan densa).
- **Análisis de Resultados:** Conclusiones basadas en los datos.
- **Análisis de Rendimiento:** Sección dedicada a las métricas del clúster (punto 2.C).

2. Repositorio de Código

Enlace al repositorio (GitHub/GitLab) estructurado que incluya:

- Scripts de ingesta y procesamiento.
- Consultas finales.
- Archivo `README.md` con instrucciones claras para desplegar y ejecutar el proyecto + docker.

3. Exposición y Defensa

Presentación oral donde se demuestre el sistema funcionando:

- **Demostración en vivo:** Ejecución del flujo de trabajo o muestra de los resultados finales en el Dashboard interactivo.
- **Revisión de infraestructura:** Mostrar brevemente el estado de los servicios para validar la salud del clúster.