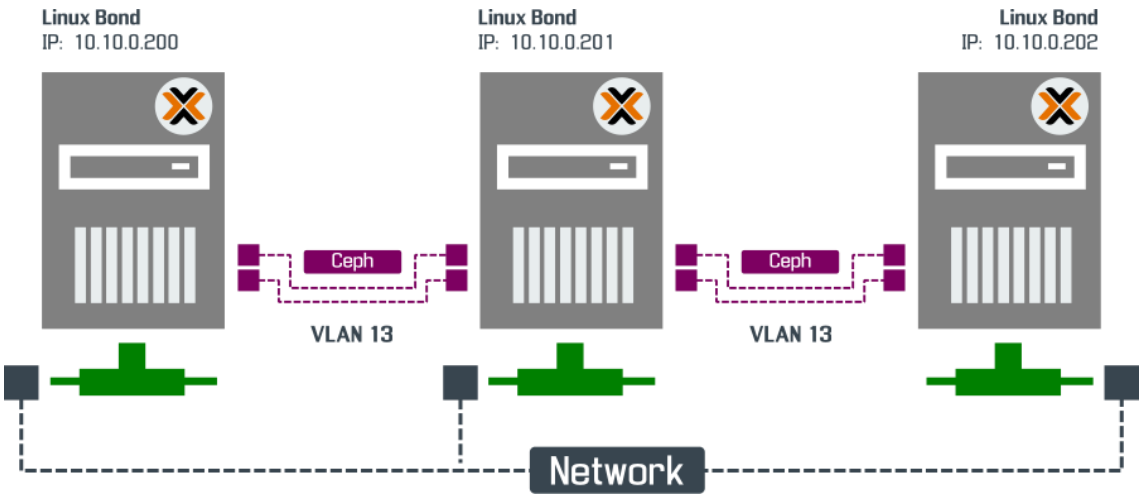
# Escuela de Ciencia de la Computación
# Examen Parcial
# CC531 A Análisis en Macrodatos 2025-II

Parte 1 [6 pts]:
Resolver en el laboratorio de clase de forma presencial 6 preguntas de teoría de forma individual, en hoja y con lapicero.

Parte 2 [14 pts]:
Se pide configurar y programar un Clúster de Hadoop e instalar Hadoop multimodo 1 maestro y 2 o más esclavos, en máquinas virtuales mínima de Ubuntu e Investigar, Desarrollar un informe, Una presentación y Exponer sobre una base de datos para su Análisis según las siguientes restricciones:




Implementar:

- Implementar un Clúster de Hadoop e instalar Hadoop multimodo **1 maestro y 2 o más esclavos**, en máquinas virtuales de Ubuntu o Windows usando virtual box u otro virtualizador a su criterio.

- Elaborar un informe en el cual:

    - **Adicional a las secciones de siempre, crear una sección inicial de implementación donde se describirá a grandes rasgos los pasos seguidos.**
    - **Adjuntando las capturas de pantalla de cada paso de configuración.**

- Implementar o usar su VM de un nodo para que comparen rendimientos en los trabajos realizados.

- Buscar un Dataset de la Plataforma Nacional de Datos Abiertos (PNDA) que no haya sido expuesto en las PC por ningún grupo (base de datos) y que sea de la **categoría Ciencia y Tecnología**.

- Tanto en el informe como en la presentación:

    - Referenciar la base de datos usada.

    - Referencias investigaciones anteriores realizadas a la base datos.

    - Indicar la fuente de dataset y la fecha de publicación, se tomará en cuenta las bases de datos más recientes.

    - Referenciar artículos que usaron su dataset si lo hubiera hecho.

- Implementar Consultas de complejidad en Hadoop
    - Hacer 3 consultas diferentes sobre promedio, la media y la desviación de un campo (Columna) a selección.
    - Hacer 2 consultas con decimales que implique al menos 3 mapreduce. Cada consulta por separado debe tener los 3 Mapreduce enlazados.
    - Hacer 2 consultas con campos decimales que implique el uso de 1 modelo de regresión y 1 modelo de clasificación. Cada consulta debe ser independiente de la otra.
- Correr su código en el clúster implementado y en la versión de un solo nodo (El que implementaron para la PC1).
- Implementar una tabla de comparación que mida: Los tiempos de ejecución, el speedup, uso promedio de CPU y memoria. La tabla y explicación de la razón de las salidas debe estar en el informe y en la presentación.
- Adicionar en el informe la imagen del monitor de recursos de la ejecución en paralelo (ejemplo htop) de memoria, de cada procesador, y tiempos de ejecución de cada caso de los nodos en la red, para observar el paralelismo de los procesos.
- En el informe y presentación para mantener la organización por cada consulta se debe mencionar:
    - Especificar el tipo de pregunta o grupo de pregunta.
    - Mencionar la consulta en el Dataset a la cual responde. Ejem: "Cual es el promedio de marcas usadas…"
    - Explicar la solución y resultados.
- Agregar al informe y presentación los inputs y outputs que pudiera tener sus consultas.
- Con los resultados obtenidos de las consultas implementar gráficos usando Power BI y adjuntarlos a su informe y presentación.
- Durante la exposición se debe mostrar la ejecución del clúster implementado distribuido y el de un solo nodo.

Donde se tenga en cuenta de:

- Tomar como base las prácticas de clase y las versiones usadas en ella de ser posible.
- Explorar y analizar la información.
- Interpretación de los resultados.
- Para la segunda parte serán en grupos de 1 o 2 personas como máximo y en el segundo caso, las dos personas deben estar presentes al momento de la exposición sino no se considerará su evaluación por más que coloquen como miembro en su presentación.
- Los grupos para el trabajo serán de 1 o 2 personas. Los grupos de 3 harán 50% más consultas. No esta permitido implementar el clúster sobre otra tecnología, de lo contrario se restarán los respectivos puntos por cada pregunta.

Comprimir todos los archivos en un archivo:

- Utilizar las plantillas mostradas en la PC1.
- Adjuntar los códigos con extensión java y también jar.
- Adjuntar el Informe con la plantilla.
- Adjuntar la presentación de su proyecto
- Entregar todo en un zip a través del aula virtual UNI hasta la fecha límite. (cada integrante del grupo debe subirlo por separado)