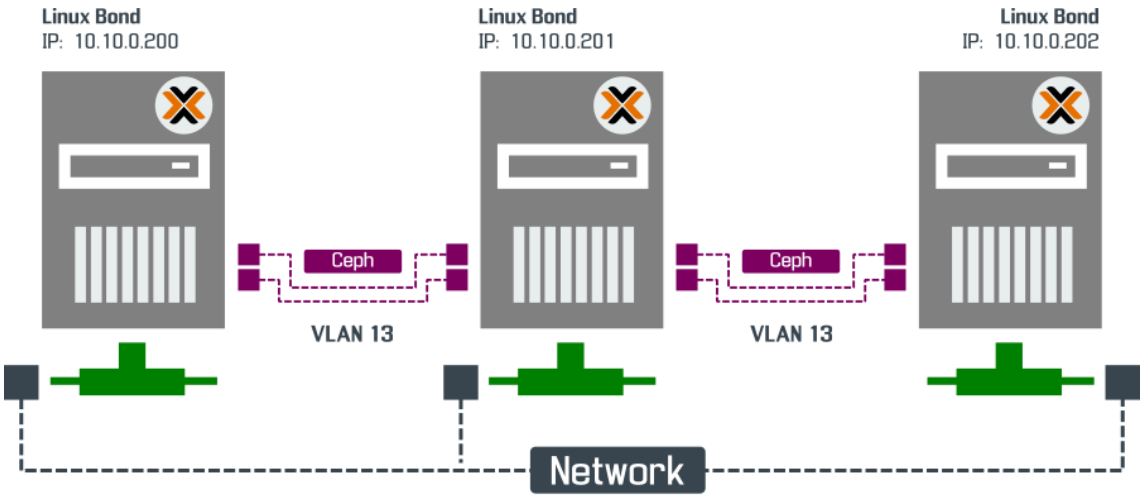
Escuela de Ciencia de la Computación
Parcial 2024-I
CC531 A Análisis en Macrodatos 2024-I

Se pide programar en un Cluster de Hadoop e instalar Hadoop multimodo 1 maestro y 2 o más esclavos, en máquinas virtuales mínima de Ubuntu y realizar un informe.



- Buscar un Dataset (base de datos).
- REFERENCIAR LA BASE DE DATOS USADA.
- REFERENCIAS INVESTIGACIONES ANTERIORES REALIZADAS A LA BASE DATOS.
- COMPARAR EL APORTE DE SUS ANÁLISIS CON LAS OTRAS INVESTIGACIONES ANTERIORES.
- 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,
- 2 consultas de complejidad en Hadoop
- 2 consultas en Hadoop donde se resuelvan con los siguientes algoritmos: linear regression, logistic regression, random forest, k-nearest neighbor, k-means clustering, naïve bayes.
- 1 consultas en Hadoop donde se resuelvan con los siguientes algoritmos: decision trees, support vector machine, redes neuronales, los grupos mayores a 2 agregara un algoritmo adicional por persona de este grupo.
- Con la pregunta o descripción y respuesta o código
- Adicionar en el informe la imagen del monitor de recursos (ejm 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.
- Adjuntar solamente los archivos con extensión java de cada consulta.
- Adjuntar los archivos con extensión jar de cada consulta.

- Adjuntar los inputs y outputs que pudiera tener sus consultas.
- Adicionar en el informe la imagen el monitor de Hadoop y de los namenode y datanodes.
- Se consultará una pregunta en su cluster o en otro cluster de clase.
- Informe.
- Presentación.
- Ejercicios propuestos en el aula:
    - El día de la evaluación a primera hora se les dejará un ejercicio por integrante, se tomará en cuenta el tiempo.
    - Para lo cual tienen que tener preparado la configuración necesaria (se tomara en cuenta, el cluster en Virtual Box).

Donde se tenga en cuenta de:

- Tomar como base las prácticas de clase.
- Explorar y analizar la información.
- Interpretación de los resultados.
- La base de datos tiene que ser publicada recientemente e indicar cuando y donde se publicó.
- Presentar el Informe y su Presentación.

Comprimir todos los archivos en un archivo:

- Utilizar la Plantilla en un archivo PDF:
    - Para describir detalles de la base de datos, el rubro de la organización de la base de datos, describir la teoría que implica los datos, que características tiene la base de datos, campos detalles, reglas de los datos, que información usted cree, que se podría obtener base de datos.
    - Donde se desarrollará y describirá el estudio de las 6 consultas de complejidad con el mayor número de campos, con sus descripción o preguntas.
    - Se tendrá en cuenta las consultas que usen varias map y varios reduce.
    - Usar el mayor número de tipo de datos y comandos de las librerías Hadoop.

No usar bases de datos relacionadas a:

- Abandono en el sector bancario, Riesgo Financiero, Enfermedades cardiovasculares (o enfermedades cardiacas), Base de datos de información de LinkedIn, Incidentes delictivos, Personas afectadas por Covid, Estilos de vestir, Casos de incendio, Pacientes, Hospitales, Vinos, Cáncer del pulmón, Venta de Casas, Spam, Wallmart, Obesidad, Recursos Humanos, Accidentes de Tránsito, Venta de Video Juego y de Laptops.

Parte 1 de la Evaluación.