UNIVERSIDAD SAN FRANCISCO DE QUITO

Informe de Trabajo en Grupo: Proyecto MapReduce para Data Mining

Desarrollado por:

- **Johana Duchi**
- **Daniela Jiménez**
- **Alex Pérez (Líder del equipo)**

Introducción

En este proyecto f se utilizó el modelo MapReduce para el procesamiento de grandes volúmenes de datos. El objetivo del proyecto fue implementar un sistema distribuido capaz de procesar un archivo de texto de aproximadamente 1GB, dividiéndolo en fragmentos más pequeños y procesando cada fragmento en paralelo, ignorando palabras vacías (stopwords) y generando un conteo de palabras. Además, se incorporó el manejo de fallos en nodos de procesamiento en las fases de Map, Shuffle y Reduce, así como en el nodo Final Reduce.

Desarrollo del Proyecto

El líder de grupo distribuyó las tareas de manera equitativa, cada miembro contribuyó de manera significativa al desarrollo del sistema MapReduce. A continuación, se describe la participación de cada integrante en el proyecto:

Johana Duchi - Fase Inicial y Fase Map

- **Responsabilidades:**
    - Se encargó de la fase inicial del proyecto, que incluyó la división del archivo de entrada de 1GB en fragmentos más pequeños de máximo 32MB. Utilizó un RandomAccessFile para leer el archivo y dividirlo en partes.
    - Ayudó a implementar el proceso de lectura de los fragmentos, asegurándose de que las palabras vacías (stopwords) fueran ignoradas para obtener un conteo preciso de palabras.
    - Desarrolló la fase Map del sistema, encargándose de procesar los fragmentos y generar pares clave-valor con la palabra y su frecuencia. Además, trabajó en la reasignación de fragmentos en caso de fallo en los nodos Map.
    - Colaboró con Alex Pérez en la implementación de los errores inducidos en los nodos Map, asegurando que el sistema pudiera reiniciar los nodos en caso de fallo.

Daniela Jiménez - Fase Shuffle y Fase Reduce

- **Responsabilidades:**

- Daniela fue responsable de la fase Shuffle del proceso MapReduce, que involucraba la reorganización de los pares clave-valor generados en la fase Map. Implementó la lógica para combinar los resultados de diferentes nodos Map en grupos de palabras únicas, facilitando el proceso de reducción.
- Desarrolló la fase Reduce, encargándose de sumar las frecuencias de palabras generadas en la fase Shuffle. Implementó el manejo de fallos en los nodos Reduce, permitiendo que el sistema reiniciara los nodos y reintentara la operación en caso de fallo.
- Colaboró en el desarrollo de la fase de errores, especialmente en la fase de Shuffle y Reduce, junto con Johana y Alex. También se encargó de probar el sistema para garantizar que funcionara correctamente en situaciones con y sin fallos.

Alex Pérez - Líder del Proyecto

- **Responsabilidades:**
  - Lideró el desarrollo del proyecto y la distribución de tareas entre los miembros del equipo.
  - Se encargó de la integración del código y de la gestión del sistema de fallos en los nodos.
  - Desarrolló el manejo de los errores y fallos en las fases Map, Shuffle, Reduce y Final Reduce, asegurando que el sistema pudiera reiniciar los nodos en caso de fallos.
  - Implementó los mecanismos de paralelización utilizando hilos (Threads) en Java para procesar los datos en paralelo.
  - Coordinó la combinación de resultados en la fase Final Reduce, encargándose de que los resultados finales de ambos MapReduce se unieran correctamente en un solo archivo de salida.

Implementación de Fallos

Los tres miembros del equipo desarrollaron la implementación de fallos inducidos. Se agregaron errores intencionales en las fases de Map, Shuffle, Reduce y en el nodo Final Reduce, y se implementó un sistema de reinicio de nodos para garantizar la capacidad de continuidad del sistema. Cada integrante trabajó en el desarrollo de los fallos en las distintas fases:

- **Fase Map:** Si se inducía un fallo en un nodo Map, el sistema reasignaba los fragmentos a otros nodos activos y continuaba el procesamiento.
- **Fase Shuffle:** Los fallos en esta fase provocaban la reasignación de los subconjuntos a otros nodos disponibles.

- **Fase Reduce:** En caso de un fallo en los nodos Reduce, el sistema reiniciaba los nodos y reintentaba el procesamiento.
- **Nodo Final Reduce:** Se implementó un sistema de reintentos en el Nodo Final Reduce, que permitía reiniciar el nodo y volver a intentar la combinación de resultados en caso de fallos.

Calificaciones

El líder de grupo asigna las siguientes calificaciones a cada uno de los integrantes según su desempeño en las tareas asignadasÑ

Johana Duchi – 91/100

Daniela Jimenez – 95/100

Alex Pérez – 88/100