**Distributed Sorting Problem**

Team Black: 20210470 조민석, 20220110 백지은, 20220871 홍지우

1. Problem of distributed sorting (in case that some of you failed to get the idea of how distributed sorting works)

Distributed sorting은 단일 컴퓨터의 메모리나 디스크 용량을 초과하는 매우 큰 데이터를 다룬다. 문제의 핵심은 여러 컴퓨터에 저장된 key/value record 정렬을 고려하는 것이다. Distributed sorting을 처리하는 방법과 해결 가능한 과제를 아래에 서술하였다.

A. 정렬 방법

1. 모든 key/value record를 읽어 메모리에 정렬하는 데 필요한 모든 데이터를 load한다.

2. 데이터가 메모리에 load되면 적절한 정렬 알고리즘을 적용하여 key/value 쌍을 정렬한다. Quick Sort나 Merge Sort 등 시간복잡도가 빠른 알고리즘(O(nlogn))을 사용하면 좋다.

3. 정렬 후 key/value record를 출력한다.

B. 과제

1. 메모리 제약

데이터셋이 너무 커서 모든 데이터를 메모리에 넣을 수 없다. 예를 들어, 입력 데이터가 50GB이고 사용 가능한 메모리가 8GB인 경우 모든 record를 메모리로 읽을 수 없다.

=>디스크 기반 병합 정렬을 사용하여 데이터의 일부를 읽고 개별적으로 정렬한 다음 여러 단계로 병합한다. 디스크를 메모리의 확장으로 사용해 더 큰 데이터셋을 효과적으로 정렬할 수 있다.

2. 디스크 용량 제한

데이터셋 크기가 시스템에서 사용할 수 있는 총 디스크 공간을 초과한다. 입력 데이터 크기가 10TB이지만 사용 가능한 디스크 용량이 1TB인 경우 단일 시스템의 메모리나 디스크가 전체 데이터 세트를 처리할 수 없다.

- 이 문제를 해결하려면 데이터셋을 여러 시스템에 분산해야 한다. 여러 시스템에 저장된 key/value record를 정렬하고 이러한 시스템에 걸쳐 정렬된 결과를 생성한다.

3. multi-core 활용

여러 시스템 간의 배포와 시스템 내 병렬화를 모두 최적화해야 한다. 각 시스템에 있는 여러 개의 CPU 코어와 여러 개의 디스크를 효과적으로 활용해야 한다.

- 병렬 정렬/분할: 각 입력 블록은 여러 코어를 사용하여 병렬로 정렬 및 분할할 수 있다. 그러나 디스크 I/O 제약으로 인해 정렬 및 분할을 위해 고정된 수의 스레드를 할당하는 것이 가장 좋다.

- 병합 단계 최적화: 기본 접근 방식에서는 각 시스템에 단일 파티션이 할당되므로 병합 단계에서는 하나의 코어만 사용할 수 있다. 더 효율적인 접근 방식은 각 시스템에 여러 개의 연속 파티션을 할당하여 병합 중에 여러 개의 코어를 동시에 사용할 수 있도록 하는 것이다.

1. Setting up git repository

https://github.com/Puzzling613/332project

1. How to effectively communicate with team members

카카오톡 단체 채팅방을 이용해 소통하고, 주마다 GSR이나 학관1층, 오아시스 회의실 등 토의할 수 있는 공간을 이용해 진행상황에 대해 이야기를 나누며 분업을 하고, 잘 안 되는 점이나 개선할 점에 대해서 이야기한다.

1. Schedule for weekly meetings, if you plan to meet physically on a regular basis

1주차: .

2주차(중간 주, 10/28 - 11/3): .

3-5주차: 정기적인 진행 상황 회의.

6주차: 진행 상황 발표 준비(마감: 11월 18일)

7주차: 프로젝트 완료를 위한 마지막 준비.

8주차: 프로젝트 마감 및 최종 프레젠테이션(12월 8일).

1. Overall plan, including milestones (which can be updated later)

3주차: Distributed Sorting의 구조 계획을 마무리하고 구현을 시작한다.

4주차: 각자 구현하기로 한 부분에 대해서 진행상황을 공유하고 어려운 점이 있으면 이야기를 나누고 해결책을 모색한다.

5주차: 첫 번째 프로토타입을 완성한다.

6주차: 진행 상황 프레젠테이션을 준비하며 프로토타입을 개선한다.

7주차: 피드백을 기반으로 구현을 개선한다.

8주차: 최종 테스트 및 프레젠테이션을 준비한다.

1. Brainstorming ideas for exploiting ChatGPT or similar AI technologies

코딩: 부분부분 간략한 구현을 chatgpt를 통해서 코드를 짜 달라고 할 수 있다.

문서화: 코드와 개발 노트를 기반으로 프레젠테이션과 보고서 등의 도움을 받을 수 있다.

디버깅: 오류 메시지를 분석하고 가능한 해결책을 제안해 줄 수 있을 것이다.

팀 커뮤니케이션: 이해하기 쉽도록 개발 노트의 가독성을 높이거나 잘 요약해 줄 수 있다.