

CSE235

데이터베이스 시스템

(Database Systems)

Practice: 파일 I/O

담당교수: 전강욱(컴퓨터공학부)

kw.chon@koreatech.ac.kr

개요

- 기본 파일 I/O
- 실습 데이터 생성
- Join 연산 구현
- External Sort 연산 구현

파일 입출력

■ 파일 입출력 과정

- 파일 열기, 읽기/쓰기, 파일 닫기
 - 파일을 반드시 닫아야, 메모리 누수 현상이 발생하지 않음

■ 파일 열기: `fopen`

- `#include <stdio.h>`
- `FILE *fp;`
- `fp = fopen("file_name", "io_mode")`

■ 입출력 방식(*io_mode*)

- `r`: 읽기 전용
- `w`: 쓰기전용
- `a`: 추가
- `r+`: 읽기/쓰기 겸용
- `w+`: 쓰기/읽기 겸용
- `a+`: 읽기/추가 겸용

파일 입출력 (계속)

■ 파일 입력 함수

- fscanf(), fgets(), fgetc()
- 열린 파일에서 내용을 읽는 함수

■ 파일 출력 함수

- fprintf(), fputs(), fputc()
- 열린 파일에 내용을 쓰는 함수

■ 파일 닫기

- fclose()

■ 참조 사이트

- <https://en.cppreference.com/w/cpp/io/c>

파일 출력 예제 (fgets.c)

- 키보드로 문자열을 입력받아 파일로 출력하는 예제

```
#include <stdio.h>

main()
{
    FILE *fp;
    int c;
    fp = fopen("./test.txt", "w"); // 쓰기 전용으로 열기
    c = getchar(); // 키보드로부터 입력받은 문자의 ASCII 코드 반환

    while(c != '.') { // '.'가 아니면
        fputc(c, fp); // fp가 가리키는 파일에 문자 c저장
        c = getchar(); // 키보드로부터 문자를 읽어 c에 저장
    }
    fclose(fp); // 파일 닫기
}
```

파일 출력 예제 (fgets.c) (계속)

■ 실행 화면

```
hadoop@hadoop-VirtualBox:~/database_example$ gcc -o fgets fgets.c
fgets.c:3:1: warning: return type defaults to 'int' [-Wimplicit-int]
  3 | main()
    | ^~~~
hadoop@hadoop-VirtualBox:~/database_example$ ls
fgets  fgets.c
hadoop@hadoop-VirtualBox:~/database_example$ ./fgets
I write sentences
And then you could check these sentences
.
hadoop@hadoop-VirtualBox:~/database_example$ cat ./test.txt
I write sentences
And then you could check these sentences
```

파일 입출력 예제 (fcopy.c)

■ fgets.c 코드를 복사하여 파일로 출력하는 예제

```
#include <stdio.h>

main()
{
    FILE *fpr;
    FILE *fpw;
    char buffer[100]; // Data를 임시로 저장하는 버퍼
    fpr = fopen("fgets.c", "r");
    fpw = fopen("tmp.c", "w");

    if (fpr == NULL) {
        printf("[ERROR] files are not found");
        exit(0);
    }

    // fpr이 가리키는 파일에서 데이터를 읽어서 buffer에 저장(최대 길이 100)
    // 이후, fpw가 가리키는 파일에 buffer의 내용 기록
    while() {
        fputs(buffer, fpw);
    }
    fclose(fpr);
    fclose(fpw);
}
```

파일 입출력 예제 (fcopy.c) (계속)

■ 실행화면

```
hadoop@hadoop-VirtualBox:~/database_example$ ls -lt
합계 48
-rwxrwxr-x 1 hadoop hadoop  479  9월  4 14:57 fcopy.c
-rw-rw-r-- 1 hadoop hadoop  197  9월  4 14:09 tmp.c
-rwxrwxr-x 1 hadoop hadoop 16224  9월  4 14:09 fcopy
-rw-rw-r-- 1 hadoop hadoop   60  9월  4 14:05 test.txt
-rwxrwxr-x 1 hadoop hadoop 16088  9월  4 14:04 fgets
-rw-rw-r-- 1 hadoop hadoop  197  9월  4 13:58 fgets.c
hadoop@hadoop-VirtualBox:~/database_example$ diff ./fgets.c tmp.c
hadoop@hadoop-VirtualBox:~/database_example$
```


실습

■ 다음 프로그램을 작성하고, 프로그램 실행결과를 함께 제출

1. `fopen()`, `fclose()`, `fgets()`, `fputs()`를 사용하여 다음 프로그램을 작성

- 아래 내용(이전 실습 시간에 작성한 내용)을 읽어서, 학번이 짝수이면, 입력파일의 홀수 줄만을 출력 파일에 저장 (학번이 홀수일 경우, 짝수 줄만을 출력)

Computer Science is the study of computers and computational systems, encompassing both theoretical knowledge and practical applications. It deals with the design, analysis, and implementation of algorithms, data structures, and software, as well as the study of the fundamental principles that underlie computation and information processing. Computer Science has a broad scope and includes various subfields, each focusing on different aspects of computing. Some of the key areas within computer science are:

1.Algorithms and Data Structures: This area focuses on designing efficient algorithms (step-by-step instructions for solving problems) and data structures (ways to organize and store data) to optimize performance and resource usage in computer programs.

2.Programming Languages: The study of different programming languages, their syntax, semantics, and how to effectively use them to develop software.

3.Software Engineering: This involves the process of designing, developing, testing, and maintaining large software systems with an emphasis on quality, reliability, and efficiency.

실습 (계속)

2. 파일에 직각이등변삼각형을 출력하는 프로그램을 작성

- 가로 및 세로 길이는 동일하며, 아래와 같은 삼각형 모양으로 출력

실습 (계속)

3. 파일을 읽어 소문자를 모두 대문자로 변경하는 프로그램을 작성

- fopen(), fclose(), fgets(), fputs()를 사용하여 작성

Computer Science is the study of computers and computational systems, encompassing both theoretical knowledge and practical applications. It deals with the design, analysis, and implementation of algorithms, data structures, and software, as well as the study of the fundamental principles that underlie computation and information processing. Computer Science has a broad scope and includes various subfields, each focusing on different aspects of computing. Some of the key areas within computer science are:

1.Algorithms and Data Structures: This area focuses on designing efficient algorithms (step-by-step instructions for solving problems) and data structures (ways to organize and store data) to optimize performance and resource usage in computer programs.

2.Programming Languages: The study of different programming languages, their syntax, semantics, and how to effectively use them to develop software.

3.Software Engineering: This involves the process of designing, developing, testing, and maintaining large software systems with an emphasis on quality, reliability, and efficiency.

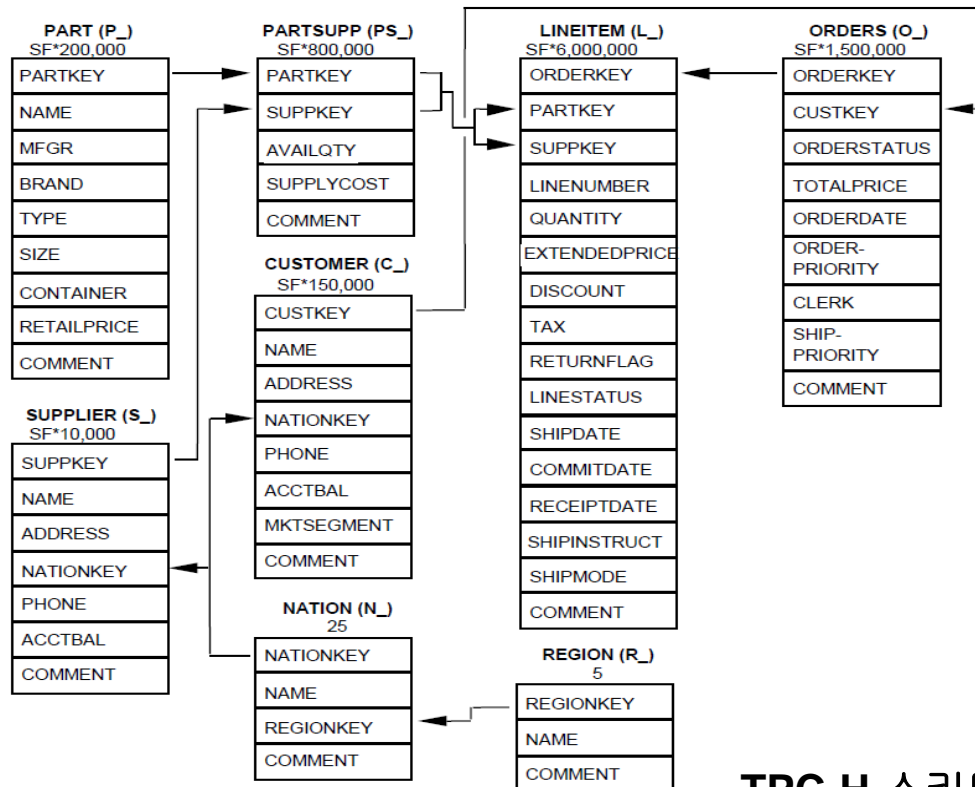
개요

- 기본 파일 I/O
- 실습 데이터 생성
- Join 연산 구현
- External Sort 연산 구현

TPC-H 벤치마크

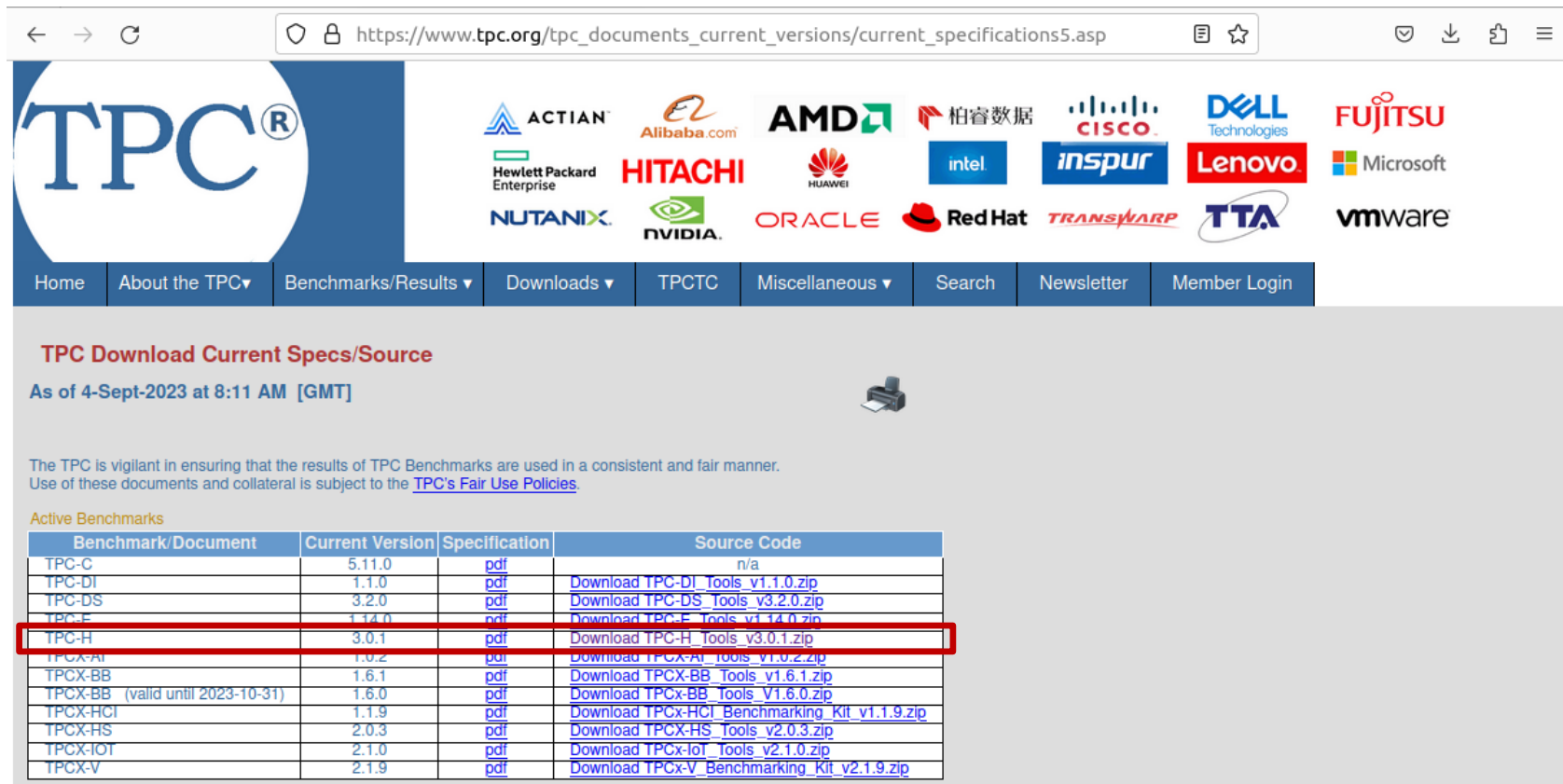
■ DBMS의 성능 측정 시 사용되는 벤치마크

- 의사 결정 용도의 시스템에 대한 성능측정
 - Business를 위한 Adhoc Query (특별한 목적을 위해서) 위주로 구성
- 8개의 테이블로 이루어진 데이터베이스에 대해 22개의 사전 정의된 Query를 제공



TPC-H 다운로드 & 데이터 생성

- TPC-H 웹사이트 접속
 - <http://www.tpc.org>
- 다운로드 페이지로 이동 후 TPC-H 클릭 (아래 그림 확인)
 - Downloads → Downloads programs and Specifications



The screenshot shows the TPC website's 'Downloads' page. The page title is 'TPC Download Current Specs/Source' as of 4-Sept-2023 at 8:11 AM [GMT]. Below the title, there is a printer icon and a disclaimer: 'The TPC is vigilant in ensuring that the results of TPC Benchmarks are used in a consistent and fair manner. Use of these documents and collateral is subject to the [TPC's Fair Use Policies](#).' The 'Active Benchmarks' section contains a table with columns: Benchmark/Document, Current Version, Specification, and Source Code. The 'TPC-H' row is highlighted with a red box.

Benchmark/Document	Current Version	Specification	Source Code
TPC-C	5.11.0	pdf	n/a
TPC-DI	1.1.0	pdf	Download TPC-DI_Tools_v1.1.0.zip
TPC-DS	3.2.0	pdf	Download TPC-DS_Tools_v3.2.0.zip
TPC-E	1.14.0	pdf	Download TPC-E_Tools_v1.14.0.zip
TPC-H	3.0.1	pdf	Download TPC-H_Tools_v3.0.1.zip
TPC-HI	1.0.2	pdf	Download TPC-HI_Tools_v1.0.2.zip
TPCX-BB	1.6.1	pdf	Download TPCX-BB_Tools_v1.6.1.zip
TPCX-BB (valid until 2023-10-31)	1.6.0	pdf	Download TPCX-BB_Tools_V1.6.0.zip
TPCX-HCI	1.1.9	pdf	Download TPCx-HCI_Benchmarking_Kit_v1.1.9.zip
TPCX-HS	2.0.3	pdf	Download TPCX-HS_Tools_v2.0.3.zip
TPCX-IOT	2.1.0	pdf	Download TPCx-IOT_Tools_v2.1.0.zip
TPCX-V	2.1.9	pdf	Download TPCx-V_Benchmarking_Kit_v2.1.9.zip

TPC-H 다운로드 & 데이터 생성(계속)

- 본인 정보입력 및 라이선스 동의 후 다운로드 클릭
- 이후 입력한 Email 주소로 다운로드 링크 수령
(다음 페이지 참조)

TPC-H Tools Download

The TPC Tools are available free of charge, however all users must agree to the licensing terms and register prior to use.
Please download and read the TPC-Tools License Agreement prior to registering for the download.

Ubuntu Software

* First Name
* Last Name
* Company / Affiliation
* Occupation
* Country
* Email
* Terms

(* Required)

Note 1: You will receive an E-mail at the address that you entered above with a link to the files to download
The TPC will not share your E-mail with anybody. - (see TPC's [Privacy Policy](#))
Submitting an invalid E-mail address will result in not being able to download the software.

로봇이 아닙니다.

reCAPTCHA
개인정보 보호 - 약관

Download Cancel

TPC-H 다운로드 & 데이터생성 (계속)

■ Email 내 링크 클릭 후, TPC-H_Tools_v3.01.zip 파일 다운로드

The image shows a two-part process for downloading TPC-H tools. The top part is an email from Info@tpc.org to chon0705@gmail.com, titled 'TPC-Tools (TPC-H) Download Confirmation'. The email contains a link to download the software, which is highlighted with a red box. The bottom part is a screenshot of the web browser at the URL https://tpc.org/tpc_documents_current_versions/download_programs/tools-download5.asp?email=chon0705@gmail.com&bm_type=TPC-H&bm_version=3.0.1&download_key=5E82BC0A%2DD479%2D4ED8%2D90C8%2D5C4982047541. The browser shows the 'TPC-H Tools Download' page, which also has a red box around the link 'TPC-H_Tools_v3.0.1.zip (Tools)'. The page includes instructions about the download process, such as the file being temporary and the link being valid for three hours.

TPC-Tools (TPC-H) Download Confirmation ▶ 받은편지함 x

Info@tpc.org 도메인: email-od.com
나에게 ▼
chon0705@gmail.com

오후 4:31 (42분 전) ☆ ↶ ⋮

Thank you for signing up to download the TPC-H Tools.

Please select the link below or copy and paste it into your web browser to download the software:

https://tpc.org/tpc_documents_current_versions/download_programs/tools-download5.asp?email=chon0705@gmail.com&bm_type=TPC-H&bm_version=3.0.1&download_key=5E82BC0A%2DD479%2D4ED8%2D90C8%2D5C4982047541

Note: A new (temporary) file is being created right now for you. Depending on the size of the file(s) this might take up to 2 minutes.

The temporary file will be available for download for about 3 hours and will be deleted then.

This link will be valid for about three hours for a single download. After that, you will have to register for a new download again.

TPC-H Tools Download
Thank you for registering to download the TPC tools software package.

• [TPC-H_Tools_v3.0.1.zip \(Tools\)](#)

Please note that some browsers block the automatic download option (e.g.: 'MS Edge'). In that case cut and paste the link from the E-mail that you have received into a different browser (e.g.: 'Google Chrome' or 'Firefox')

Depending on your network connection and the size of the file to be downloaded, it might take 30 minutes or even more for the download to finish (TPCx-V - 1.8 GB) - please be patient. Most of the downloads will finish within a few seconds.

The file can only be downloaded once. If you don't see a file to download on this screen, please register again [here](#).

If you don't see a link to download the tools you have requested, please click [here](#).

TPC-H 다운로드 & 데이터 생성 (계속)

■ make 설치

- `$ sudo apt install make`

■ TPC-H 파일 압축 해제 및 파일 생성

- `$ unzip TPC-H-Tool.zip`
 - 다운로드 받은 파일을 압축해제 하는 것이며, 압축파일 명은 다를 수 있음
- `$ cd 'TPC-H V3.0.1'`
- `$ cd dbgen`
- `$ cp makefile.suite Makefile`
- `$ vi Makefile` // Makefile 내 아래 내용 변경
 - `DATABASE = SQLSERVER`
 - `MACHINE = LINUX`
 - `WORKLOAD = TPCH`
 - `CC = gcc`
- `$ make dbgen`
- `$ time ./dbgen`

개요

- 기본 파일 I/O
- 실습 데이터 생성
- **Join 연산 구현**
- External Sort 연산 구현

조인(Join) 연산

- 조인 연산은 2개 또는 그 이상의 테이블 대상으로 질의할 때 사용됨
 - 테이블들의 특정 컬럼의 관계에 기반하여 연산을 수행

```
SELECT column_name(s)
FROM table_name1, table_name2
ON table_name1.column_name = table_name2.column_name;
```

Table : Grade

Id	Grade
1	A
2	B
3	A

Table : Student

Id	Name
1	John
2	Make
3	Deny

SELECT *
FROM Student, Grade
ON Student.id = Grade.id;



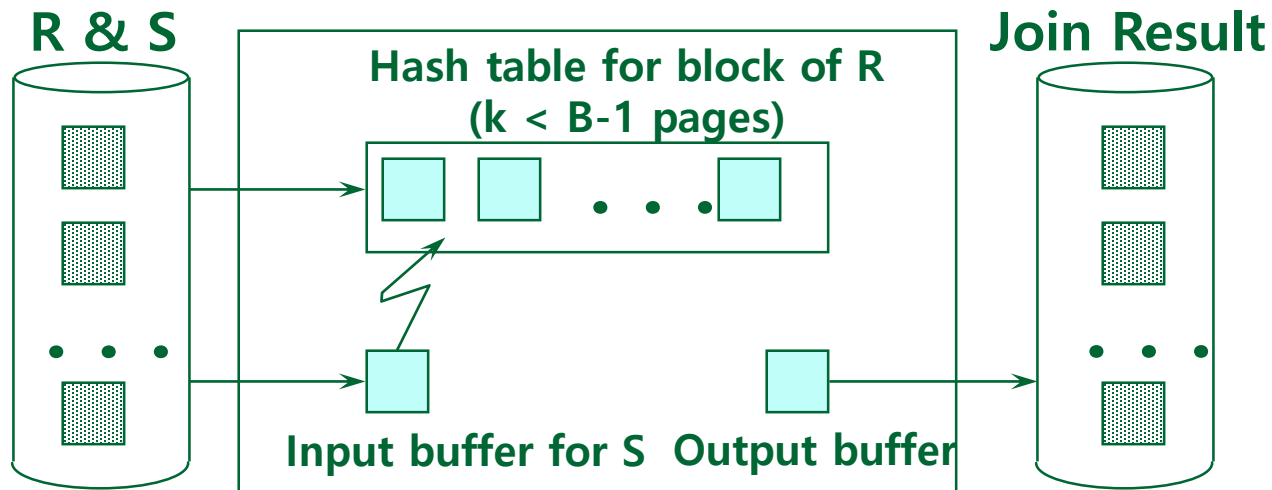
Id	Name	Grade
1	John	A
2	Make	B
3	Deny	A

Nested Loops Join

■ Simple Nested Loops Join

```
foreach tuple r in R do
  foreach tuple s in S do
    if  $r_i == s_j$  then add  $\langle r, s \rangle$  to result
```

■ Block Nested Loops Join



실습

- TPC-H 내 PART 테이블과 PARTSUPP 테이블에 대한 **Block Nested Loops Join** 연산을 C/C++로 구현
 - 입력: PART 테이블과 PARTSUPP 테이블은 파일로 디스크에 저장되어 있음
 - 연산: Join 컬럼은 PARTKEY이며 동등 조인만 구현
 - 출력: 조인의 결과는 파일로 저장
- 아래 내용 포함하는 보고서 작성 후 소스코드 전문과 함께 제출
 - 전체적인 구현 Details 설명
 - 성능 분석
 - 버퍼 크기를 조절하면서, 수행시간, 메모리 Footprint 등

개요

- 기본 파일 I/O
- 실습 데이터 생성
- Join 연산 구현
- **External Sort 연산 구현**

외부 정렬 (External Sort)

- 외부 정렬은 디스크에 저장된 파일의 정렬을 의미 (내부 정렬(internal sort)은 RAM에 있는 데이터 배열을 정렬하는 것임)
 - 외부 정렬의 주된 관심사는 디스크 접근 횟수를 줄이는 것임
 - 데이터가 메인 메모리에 저장하기에 너무 큰 경우에 주로 사용됨
 - 예: 대용량 데이터베이스, 거대 3D 모델 기반의 그래픽스 응용 등에 활용
- 외부 메모리 병합 정렬(external memory merge-sort)
 - 정렬할 파일을 처음부터 훑으면서 주 메모리 크기 정도로 블록을 나눈 후 이러한 블록을 정렬
 - 이후 정렬된 블록을 병합

실습

- **TPC-H의 lineitem.tbl 테이블의 레코드를 고정크기 레코드로 변경**
 - 임의의 컬럼 기준으로 정렬 가능하도록 프로그램 할 것
- **고정 크기 레코드 파일을 디스크에 저장하고, 주어진 주 메모리 양을 사용하여 정렬하기**
 - 메인 메모리 내 허용되는 레코드 개수 조절 가능하도록 프로그램 할 것
- **구현 고려사항**
 - I/O stream은 디스크에서 B 크기의 블록을 한번에 읽거나 쓸 수 있도록 여러 개의 크기 B의 버퍼를 메모리에 유지해야 함
 - 주어진 메모리 크기 M을 초과하지 않도록 해야 함
- **아래 내용 포함하는 보고서 작성 후 소스코드 전문과 함께 제출**
 - 전체적인 구현 Details 설명
 - 성능 분석
 - 버퍼 크기를 조절하면서, 수행시간, 메모리 Footprint 등

감사합니다.

Contact: kw.chon@koreatech.ac.kr