

CSE235

database system

(Database Systems):

Implementing database operations

Professor in charge: Jeon Kang-wook (Department of Computer Engineering)

kw.chon@koreatech.ac.kr

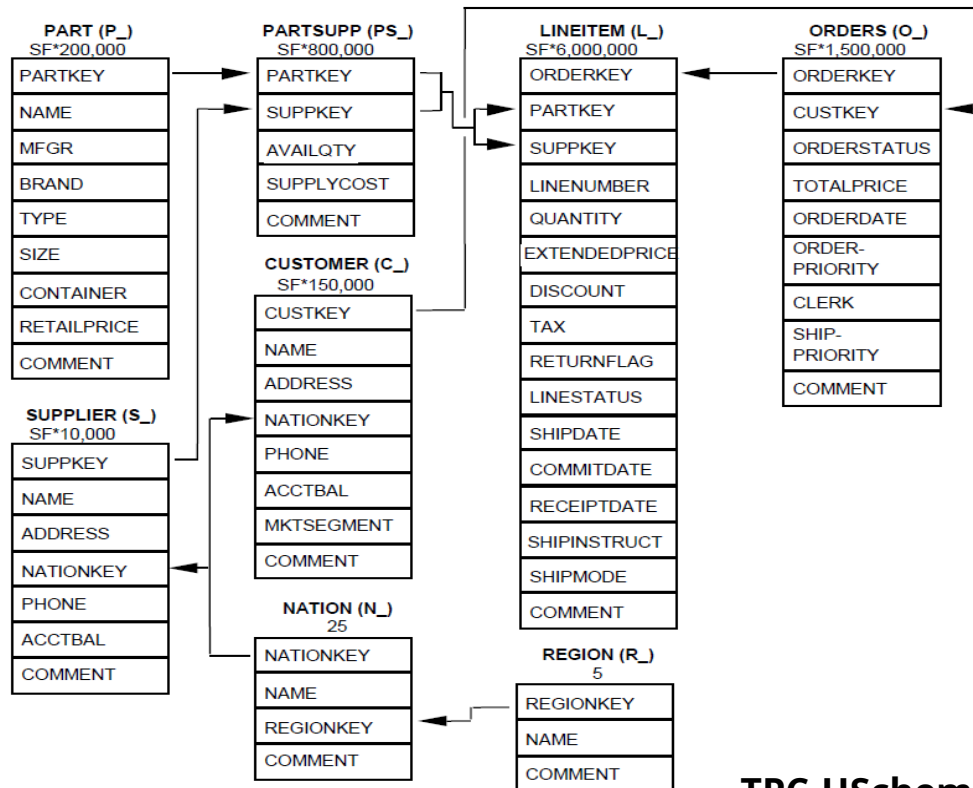
outline

- **Generate practice data**
- Implementing Join Operations
- External Sort Operation Implementation
- Project Overview

TPC-H Benchmark

■ Benchmarks used to measure the performance of DBMS

- Performance measurement for systems for decision making purposes
 - Adhoc Query for Business (for special purposes)
- Provides 22 predefined queries for a database consisting of 8 tables.



TPC-H Download & Data Generation

- Access the TPC-H webpage
 - <http://www.tpc.org>
- Go to the download page and click TPC-H (see image below)
 - Downloads → Downloads programs and specifications

The screenshot shows the TPC website's download page. The header features the TPC logo and a row of partner logos including ACTIAN, Alibaba.com, AMD, 柏睿数据, CISCO, DELL Technologies, FUJITSU, Hewlett Packard Enterprise, HITACHI, HUAWEI, intel, inspur, Lenovo, Microsoft, NUTANIX, NVIDIA, ORACLE, Red Hat, TRANSWARP, TTA, and vmware. The navigation bar includes links for Home, About the TPC, Benchmarks/Results, Downloads, TPCTC, Miscellaneous, Search, Newsletter, and Member Login. The main content area is titled "TPC Download Current Specs/Source" and shows the date "As of 4-Sept-2023 at 8:11 AM [GMT]". A disclaimer states: "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](#)." Below this, a section titled "Active Benchmarks" contains a table with four columns: Benchmark/Document, Current Version, Specification, and Source Code. The TPC-H row is highlighted with a red rectangle.

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 Download & Data Generation (Continued)

- After entering your information and agreeing to the license, click Download
- After that, you will receive a download link to the email address you entered (see next page)

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

☒ I have read and agree to the [TPC End User License Agreement](#) - (.txt file).

TPC-H Download & Data Generation (Continued)

- After clicking the link in the email, download the TPC-H_Tools_v3.01.zip file.

The screenshot shows an email from Info@tpc.org to chon0705@gmail.com. The email subject is "TPC-Tools (TPC-H) Download Confirmation". The body of the email says: "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:" followed by a 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. Below the link, there is a note: "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."

The browser window shows 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 page title is "TPC-H Tools Download". The page content says: "Thank you for registering to download the TPC tools software package." followed by a list of download links: [TPC-H_Tools_v3.0.1.zip \(Tools\)](#). Below the list, there is a note: "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 dont 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 Download & Data Generation (Continued)

■ make install

- ❑ \$sudo apt install make

■ Unpacking TPC-H files and creating files

- ❑ \$unzip TPC-H-Tool.zip
 - Unzip the downloaded file, and the compressed file name may be different.
- ❑ \$cd 'TPC-H V3.0.1'
- ❑ \$cd dbgen
- ❑ \$cp makefile.suite Makefile
- ❑ \$vi Makefile// Change the following in Makefile
 - DATABASE = SQLSERVER
 - MACHINE = LINUX
 - WORKLOAD = TPCH
 - CC = gcc
- ❑ \$make dbgen
- ❑ \$time ./dbgen

outline

- Generate practice data
- **Implementing Join Operations**
- External Sort Operation Implementation
- Project Overview

Join operation

- The join operation is used when querying two or more tables.

□ Perform operations based on the relationships between specific columns in tables.

```
SELECTcolumn_name(s)  
FROMtable_name1, table_name2  
ONtable_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

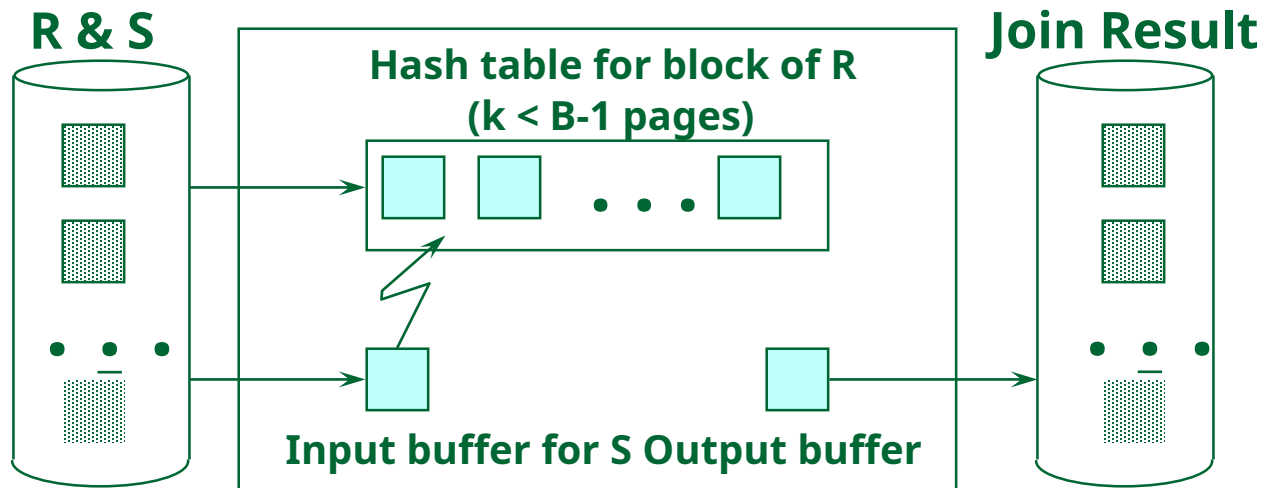
■ Simple Nested Loops Join

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



Practice

- **About the PART table and PARTSUPP table in TPC-H** **Block**
Nested Loops Join **Implementing operations in C/C++**
 - **input:** The PART table and PARTSUPP table are stored on disk as files.
 - **calculation:** The Join column is PARTKEY and only equi-join is implemented.
 - **output of power:** The results of the join are saved to a file.

- **After writing a report including the contents below, submit it along with the source code.**
 - Description of overall implementation details
 - Performance Analysis
 - Adjusting the buffer size, execution time, memory footprint, etc.

outline

- Generate practice data
- Implementing Join Operations
- **External Sort Operation Implementation**
- Project Overview

External Sort

- **External sorting refers to sorting files stored on disk (internal sorting refers to sorting the data array in RAM)**

- The main concern with external sorting is to reduce the number of disk accesses.
- Mainly used when data is too large to be stored in main memory.
- Examples: Used in large databases, graphics applications based on huge 3D models, etc.

- **External memory merge-sort**

- It sorts the file to be sorted by dividing it into blocks of the size of the main memory from the beginning and sorting these blocks.
- Merge the sorted blocks afterwards

Practice

- **Change records in lineitem.tbl table of TPC-H to fixed size records**

- ☐ I will program it so that it can be sorted by any column.

- **Store fixed-size record files on disk and sort them using a given amount of main memory.**

- ☐ Program to allow control over the number of records allowed in main memory.

- **Implementation Considerations**

- ☐ An I/O stream must maintain multiple buffers of size B in memory so that it can read or write blocks of size B from disk at a time.
- ☐ Must not exceed the given memory size M

- **After writing a report including the contents below, submit it along with the source code.**

- ☐ Description of overall implementation details

- ☐ Performance Analysis

- Adjusting the buffer size, execution time, memory footprint, etc.

outline

- Generate practice data
- Implementing Join Operations
- External Sort Operation Implementation
- **Project Overview**

Development Contents

- **Implementation of operations to save and read practice data (TCP-H) to disk**
 - Fixed-size blocks are implemented to contain multiple variable-length records.
- **Implementing Join Operation or External Sort Operation**
- **Optimization (efforts to make calculations faster) is a plus**
 - Ability to process multiple blocks simultaneously, parallelization, etc.
- **However, development using C/C++**

Presentation and Submissions

- **Presentation: Prepare within 10 minutes including Q&A**

- **Report: Prepared including the following (submitted by December 20, 2024)**
 - Implementation details
 - A way to store data on disk
 - Optimization Details
 - How the implementation works (can be demonstrated outside of class hours)
 - Performance Evaluation

thank you

Contact: kw.chon@koreatech.ac.kr