CSE235
database system
(Database Systems):

Implementing database operations

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

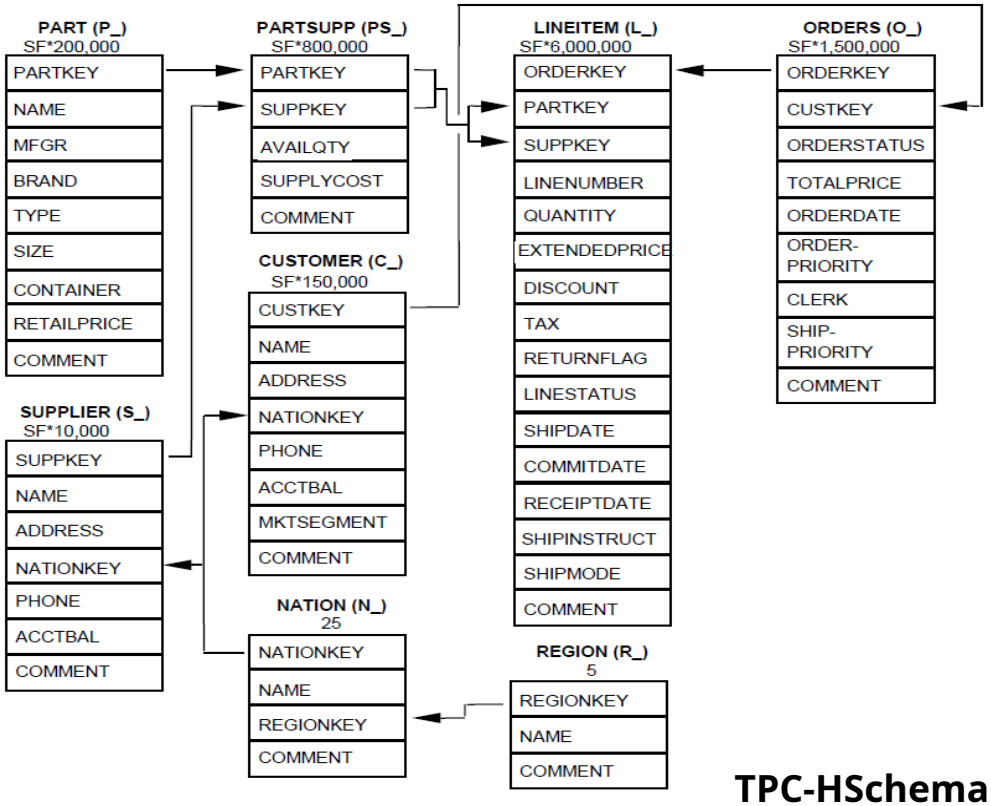
kw.chon@koreatech.ac.kr

outline

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

TPC-H Benchmark

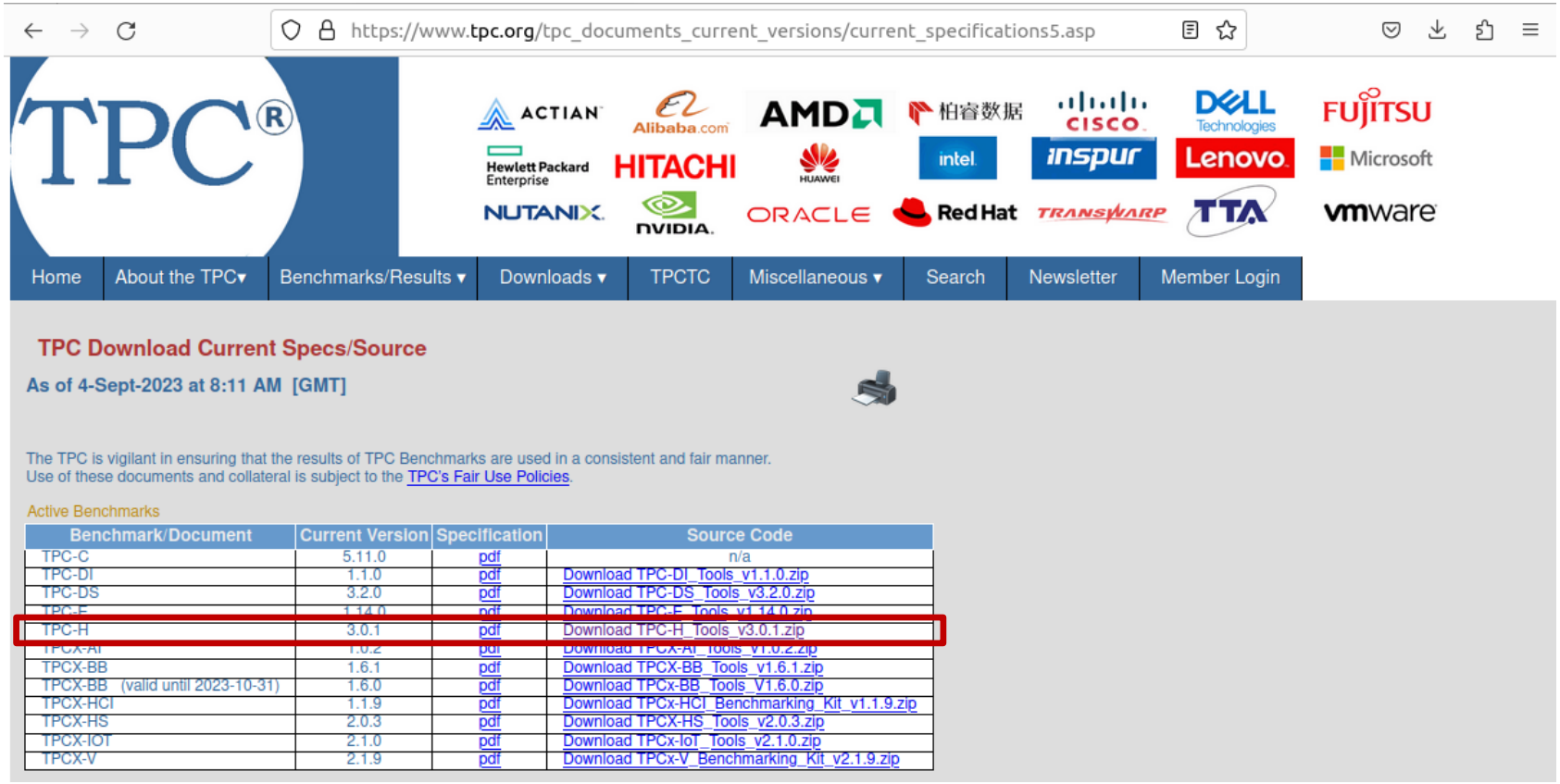
- **Benchmarks used to measure the performance of DBMS**
  - Performance measurement for systems for decision making purposes
    - Adhoc Query for Business (for special purposes)
  - Provides 22 predefined queries for a database consisting of 8 tables.



TPC-HSchema.

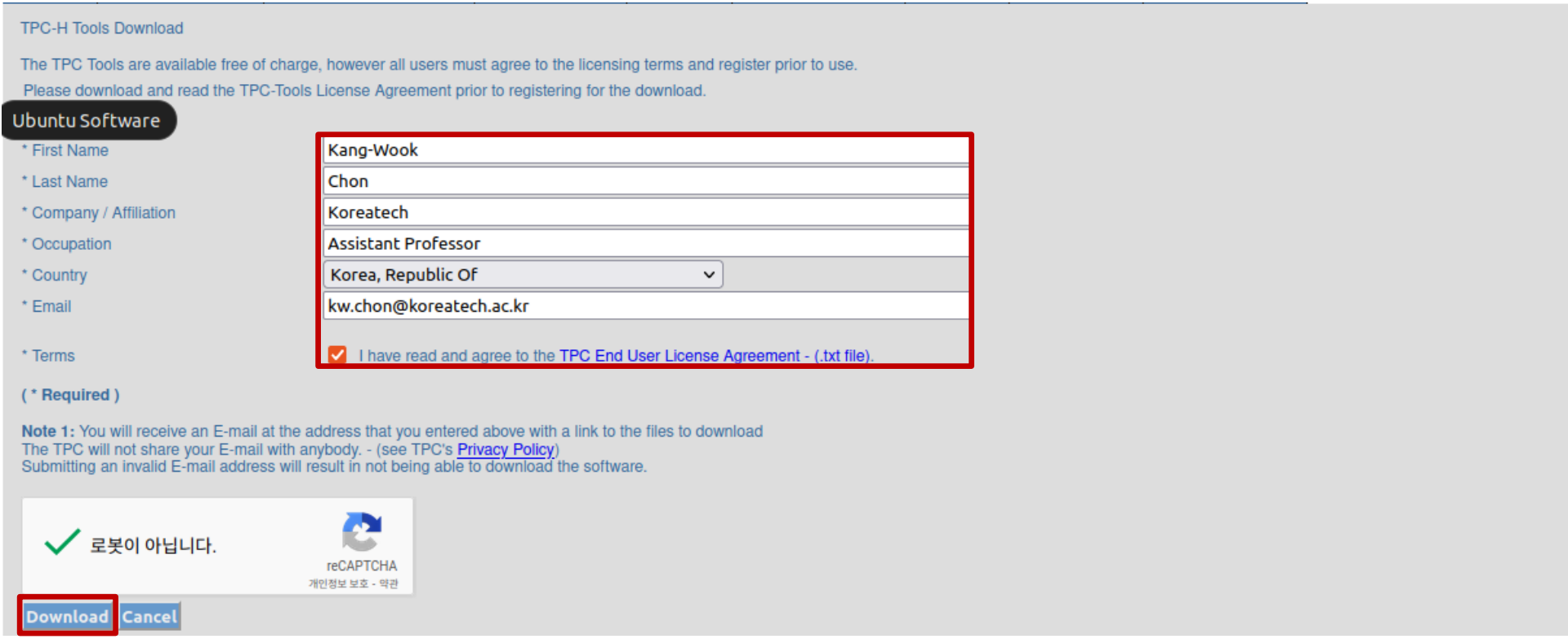
TPC-H Download & Data Generation

- **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



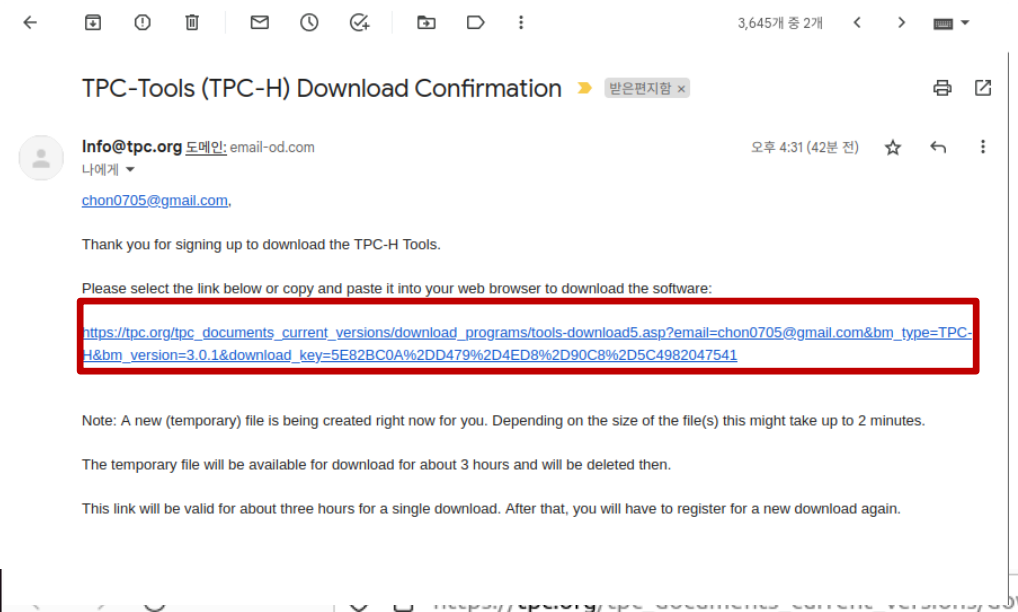
TPC-H Download & Data Generation (Continued)

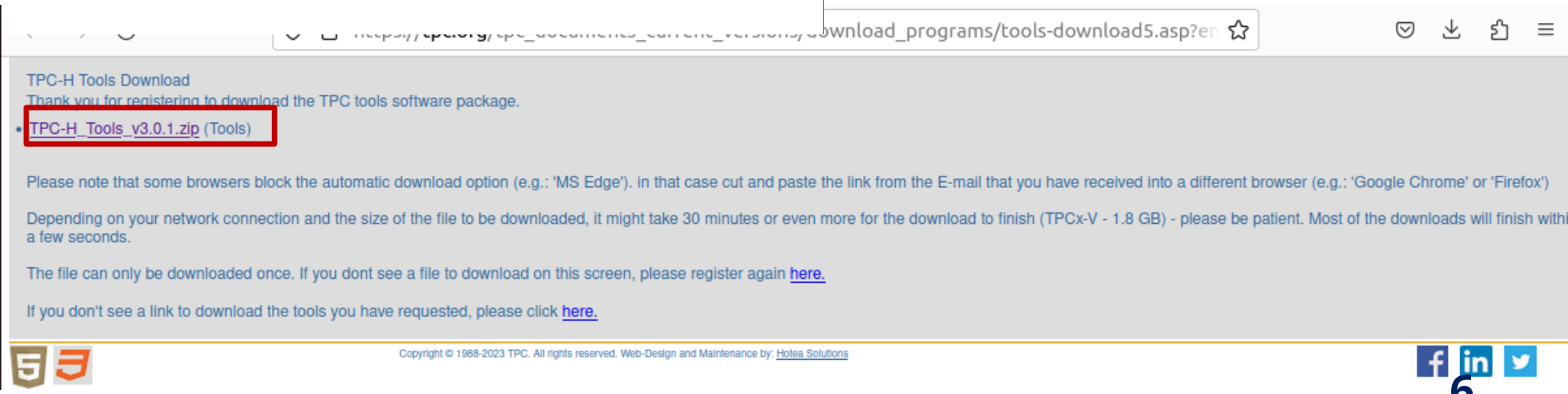
- After entering your information and agreeing to the license, click Download
- After that, you will receive a download link to the email address you entered (see next page)



TPC-H Download & Data Generation (Continued)

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





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;*

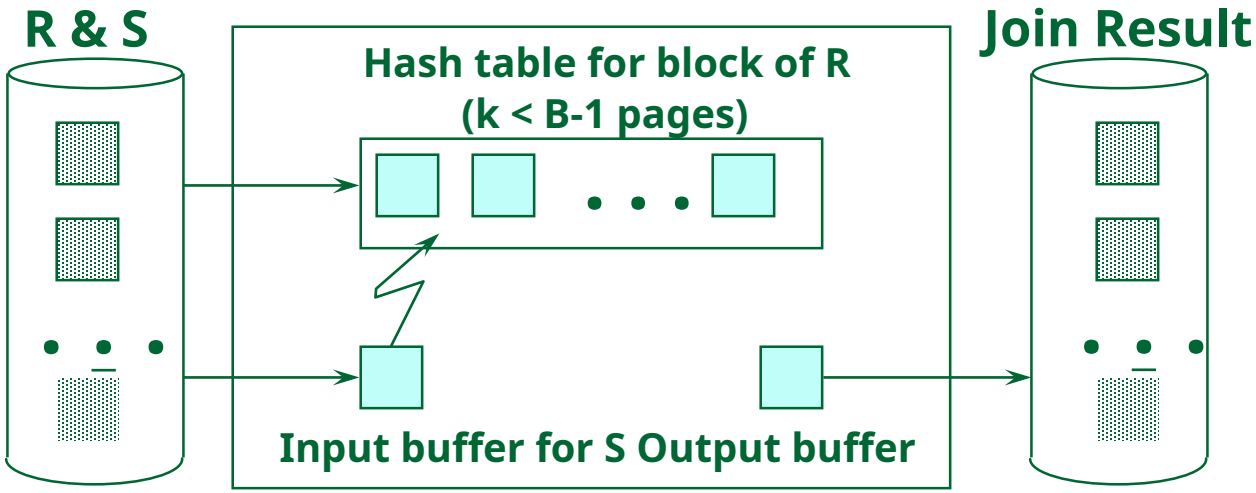
Id	Name	Grade
1	John	A
2	Make	B
3	Deny	A

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 <r, s> to result

- **Block Nested Loops Join**



*** RclassSis a relation.**

Practice

- **About the PART table and PARTSUPP table in TPC-H Block Nested Loops JoinImplementing 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