แผนการสอนประจำบทเรียน

รายชื่ออาจารย์ผู้จัดทำ ผู้ช่วยศาสตราจารย์ ณัฏฐพร พิมพายน และ สุนีย์ พงษ์พินิจภิญโญ

รายละเอียดของเนื้อหา

ตอนที่ 8.1 แนวคิดของกระบวนการสอบถามข้อมูล

เรื่องที่ 8.1.1 วัตถุประสงค์ของการทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด (query optimization)

เรื่องที่ 8.1.2 ความสัมพันธ์ระหว่างภาษาสอบถามเชิงโครงสร้างกับกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

ตอนที่ 8.2 การทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

เรื่องที่ 8.2.1 ขั้นตอนของการทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

ตอนที่ 8.3 การประเมินค่าใช้จ่ายที่เกิดขึ้นของการทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

เรื่องที่ 8.3.1 การกำหนดฟังก์ชันของค่าใช้จ่าย

เรื่องที่ 8.3.2 การเปรียบเทียบเทคนิคที่นำมาประยุกต์ใช้ที่ก่อให้เกิดค่าใช้จ่ายที่ต่ำที่สุดในการทำที่ก่อให้เกิดค่าใช้จ่ายที่ต่ำที่สุดและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

แนวคิด

1. ภาษาสอบถามเชิงโครงสร้าง (SQL) นั้นผู้ใช้จะระบุต้องการข้อมูลที่ต้องการแทนที่ผู้ใช้จะต้องคำนึงถึงวิธีว่าจะดึงข้อมูลที่ต้องการมาได้อย่างไร เป็นหน้าที่ของระบบการจัดการฐานข้อมูลจะเป็นตัวจัดการในเรื่องเกี่ยวกับการดึงข้อมูลที่ผู้ใช้ต้องการและมีประสิทธิภาพสูงสุดในการสืบค้นข้อมูล ซึ่งจะเป็นการป้องกันไม่ให้ผู้ใช้ต้องมายุ่งเกี่ยวกับวิธีการดึงข้อมูลซึ่งอาจจะเป็นวิธีที่ไม่มีประสิทธิภาพ นอกจากนั้นจะทำให้ระบบจัดการฐานข้อมูลยังสามารถควบคุมการทำงานทั้งระบบได้
2. การแปลงภาษาเอสคิวแอลที่ถูกต้องให้เป็นรูปแบบภายใน (internal representation) มีได้หลายรูปแบบ ในการเลือกรูปแบบภายในที่ก่อให้เกิดค่าใช้จ่ายต่ำสุดและได้คำตอบที่ดีที่สุดนั้นเป็นการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด (query optimization) ซึ่งอยู่ในกระบวนการสอบถามข้อมูล

วัตถุประสงค์

เมื่อศึกษาบทเรียนที่จบแล้ว นักศึกษาสามารถ

1. เข้าใจถึงแนวความคิดกระบวนการสอบถามข้อมูลซึ่งเป็นกระบวนการในการสร้างภาษาสอบถามเชิงโครงสร้างซึ่งเป็นภาษาสอบถามระดับสูงให้อยู่ในรูปของรูปแบบภายในในระดับภายใน (internal view) ของฐานข้อมูล
2. บอกถึงการทำงานของกระบวนการสอบถามข้อมูลที่เป็นส่วนหนึ่งของระบบการจัดการฐานข้อมูล

กิจกรรมการเรียนการสอน

กิจกรรมที่นักศึกษาต้องทำสำหรับการเรียนการสอน ได้แก่

1. ศึกษาเอกสารการสอน
2. ทำกิจกรรมในแบบฝึกปฏิบัติบทเรียนที่ 8
3. ทำแบบประเมินผลของบทเรียนที่ 8

สื่อการสอน

เอกสารการสอนของชุดวิชา

1. แบบฝึกปฏิบัติ
2. บทความ/ข้อมูลทางคอมพิวเตอร์
3. การให้คำปรึกษาทางโทรศัพท์
4. CD-ROM
5. Homepage ของชุดวิชาผ่านทางอินเตอร์เน็ต

เอกสารประกอบการสอน

1. Fundamentals of Database Systems, by Ramez Elmasri, Shamkant B. Navathe, The Second Edition, 1994
2. Database System Concepts, by Abraham Siberschaty, Henry F.Korth, S.Sudarshan, The Third Edition, 1991

ประเมินผล

1. ประเมินผลจากแบบฝึกหัด/ทดสอบ ในแต่ละบท
2. ประเมินผลจากการสอนประจำภาคการศึกษา

ตอนที่ 8.1 กระบวนการสอบถามข้อมูล

หัวเรื่อง

เรื่องที่ 8.1.1 วัตถุประสงค์ของการทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด (query optimization)

เรื่องที่ 8.1.2 ความสัมพันธ์ระหว่างภาษาสอบถามเชิงโครงสร้างกับกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

แนวคิด

1. การสอบถามข้อมูลในฐานข้อมูลนั้นเป็นการสอบถามข้อมูลที่อยู่ในรูปของภาษาสอบถามเชิงโครงสร้าง (SQL) ซึ่งเป็นภาษาระดับสูง เมื่อผู้ใช้ต้องการสอบถามข้อมูลจากฐานข้อมูลโดยใช้ภาษาเอสคิวแอลนั้น ระบบการจัดการฐานข้อมูลจะทำการตรวจสอบไวยากรณ์และความถูกต้องตามหลักภาษาและแปลงภาษาเอสคิวแอลที่ถูกต้องให้เป็นรูปแบบภายใน (internal representation) และทำการดึงข้อมูลจากฐานข้อมูลตามที่ผู้ใช้ต้องการ ซึ่งต้องอาศัยกระบวนการสอบถามข้อมูล
2. การสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุดนั้นเป็นขั้นตอนหนึ่งในกระบวนการสอบถามข้อมูล เมื่อผู้ใช้สามารถใช้ภาษาเอสคิวแอลในการสอบถามว่าจะต้องการข้อมูลอะไร กระบวนการสอบถามข้อมูลในระบบการจัดการฐานข้อมูลจะเป็นตัวจัดการโดยที่จะเลือกแผนปฏิบัติ (query execution plan) ที่ทำให้ได้คำตอบที่ดีที่สุด

วัตถุประสงค์

หลังจากที่ศึกษาตอนที่ 8.1 แล้ว นักศึกษาสามารถ

1. เข้าใจถึงความสำคัญของกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด
2. ความสัมพันธ์ระหว่างภาษาสอบถามเชิงโครงสร้างกับกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

เรื่องที่ 8.1.1. วัตถุประสงค์ของการทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด (query optimization)

8.1.1.1 ความหมายของการประมวลผลสอบถาม

การประมวลผลสอบถาม (query processing) เป็นหน้าที่หนึ่งของระบบการจัดการฐานข้อมูล โดยระบบการจัดการฐานข้อมูลจะทำการประมวลผลสอบถามที่เกิดจากภาษาสอบถาม (query language) ได้แก่ ภาษา SQL (Structured Query Language) ภาษา QBE (Query By Example) เป็นต้น และหาคำตอบที่ดีที่สุดและถูกต้องตรงกับความต้องการให้แก่ผู้ใช้ ได้อย่างเหมาะสมและรวดเร็วและประหยัดค่าใช้จ่ายให้เกิดน้อย

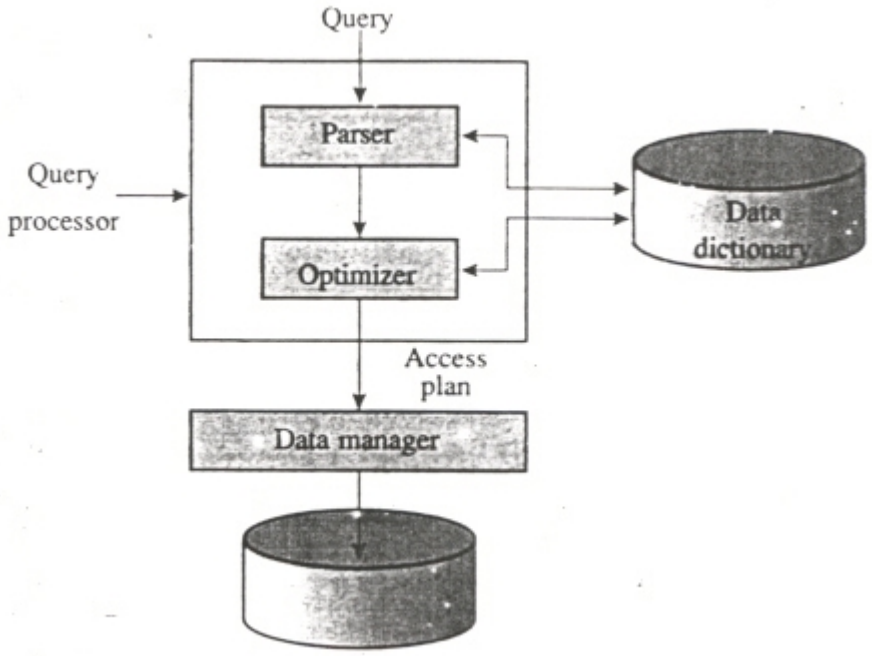
ที่สุดในส่วนที่เกี่ยวข้องกับเวลาที่ใช้ในการประมวลผลที่หน่วยประมวลผลกลาง (CPU Time) และเวลาที่ใช้ในการดึงข้อมูลจากอุปกรณ์อินพุทและเอาท์พุท (I/O time)

การประมวลผลสอบถาม (query processing) เป็นกระบวนการหนึ่งของระบบการจัดการฐานข้อมูลที่ใช้ในการเลือกแผนหรือกลยุทธ์ที่เหมาะสมในการสอบถามข้อมูลในฐานข้อมูล โดยระบบจัดการฐานข้อมูลจะมีตัวประมวลผลที่เรียกว่า ตัวประมวลผลสอบถาม (query processor) ทำหน้าที่ในการเลือกแผนหรือกลยุทธ์ที่เหมาะสมในการเข้าถึงข้อมูลเพื่อให้ได้คำตอบที่ดีที่สุดและถูกต้องตรงกับความต้องการให้แก่ผู้ใช้มากที่สุด

เรื่องที่ 8.1.2 ความสัมพันธ์ระหว่างภาษาสอบถามเชิงโครงสร้างกับกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

8.1.2.1 โครงสร้างของตัวประมวลผลสอบถาม

ตัวประมวลผลสอบถามเป็นส่วนประกอบหนึ่งของระบบการจัดการฐานข้อมูลที่ทำหน้าที่ในการเลือก



แผนหรือกลยุทธ์ที่เหมาะสม ตัวประมวลผลสอบถามมีโครงสร้างดังภาพที่ 8.1.2.1 ซึ่งมีส่วนประกอบดังนี้

ภาพที่ 8.1.2.1 โครงสร้างของตัวประมวลผลสอบถาม

8.1.2.1.1 **ส่วนตรวจสอบภาษา** (parser)

เป็นส่วนหนึ่งของตัวประมวลผลสอบถามที่ทำหน้าที่ตรวจเช็คกฎและหลักเกณฑ์ (syntax) ของคำสั่งสอบถามข้อมูลที่กำหนดไว้ในระบบฐานข้อมูล โดยจะทำการตรวจสอบข้อมูลที่อยู่ใน

พจนานุกรมข้อมูลแล้วทำการแปลคำสั่งให้อยู่ในรูปของสมการพีชคณิตแบบสัมพันธ์ (relational algebraic expression) ที่ระบบจัดการฐานข้อมูลสามารถเข้าใจได้

ตัวอย่าง ภาษาSQL เพื่อภาษาสอบถามข้อมูลต้องการสอบถามข้อมูลจากตาราง COURSE โดยทำการค้นหาข้อมูลที่มีรหัสวิชา (Course#) เป็น COMP353 ที่มี Smith เป็นผู้สอนมีดังนี้

```
select Course#
from COURSE
where Course# = COMP353 and
      Insturctor = 'Smith';
```

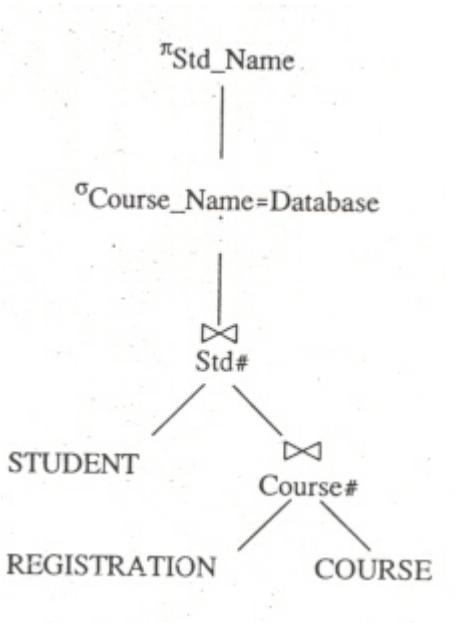
จากคำสั่งภาษา SQL ดังกล่าวข้างต้น ส่วนตรวจสอบภาษา (parser) จะทำการแปลคำสั่งในการสอบถามจากภาษา SQL ให้อยู่ในรูปของสมการพีชคณิตแบบสัมพันธ์ (relational algebraic expression) ได้ดังนี้

$$\sigma_{Course\#} = COMP353 \wedge (Instructor = {}'Smith'(COURSE))$$

8.1.2.1.2 **ออฟติไมเซอร์** (optimizer)

เป็นส่วนหนึ่งของตัวประมวลผลสอบถามที่ทำหน้าที่ในการกลยุทธ์หรือวิธีการที่เหมาะสมเพื่อให้การเข้าถึงข้อมูลประหยัดเวลาที่สุด เมื่อคำสั่งในภาษาสอบถามผ่านตัวตรวจสอบภาษา (parser) แล้วจะได้คำสั่งในการสอบถามที่อยู่ในรูปของสมการพีชคณิตแบบสัมพันธ์ (relational algebraic expression) ซึ่งสามารถประมวลผลได้หลายวิธีด้วยเทคนิคต่าง ๆ ในการหาทางเลือกที่เหมาะสม (optimization) ออฟติไมเซอร์จะทำการเลือกการเข้าถึงข้อมูล (aceess plan) โดยการเลือกแผนการเข้าถึงข้อมูลซึ่งจะต้องคำนึงถึงประเด็นต่อไปนี้

- การลดปริมาณการดึงหรือเรียกใช้ข้อมูล
- การลดขนาดของโครงสร้างข้อมูลโดยพิจารณาใช้เทคนิคการลดการทำงานระหว่างการดึงข้อมูล
- เพิ่มความสามารถในการดึงข้อมูลด้วยการเรียงลำดับไว้ก่อน การจัดเรียงลำดับใหม่อาจจะช่วยลดการทำงานบางอย่างไร เพื่อช่วยให้การประมวลผลสอบถามสามารถทำงานได้ดีขึ้น ออฟติไมเซอร์โดยอาจสร้างสมการพีชคณิตแบบสัมพันธ์ให้อยู่ในรูปแบบของโครงสร้างต้นไม้ที่อยู่ในรูปของกราฟ (operator graph) ดังภาพที่ 8.1.2.2 เป็นโอเปอร์เรเตอร์กราฟ (operator graph) ของการค้นหารายชื่อของนักศึกษาที่ลงทะเบียนเรียนวิชา Database



ภาพที่ 8.1.2.2 ตัวอย่างโอเปอร์เรชันกราฟ(operator graph)

- จัดลำดับคำสั่งเพื่อให้ขั้นตอนการทำงานลดลงได้แก่ จัดลำดับการคำสั่ง SELECT เพื่อลดการนำเข้าข้อมูลตั้งแต่เริ่มต้นดึงข้อมูลจากดิสก์
- เลือกใช้คำสั่งในการจัดกลุ่ม(PROJECT) และคำสั่งค้นหา (SELECT) ในการเลือกกลุ่มข้อมูล

กิจกรรม 8.1

การประมวลผลสอบถามคืออะไร

แนวตอบกิจกรรม 8.1

1. การประมวลผลสอบถาม เป็นกระบวนการหนึ่งของระบบการจัดการฐานข้อมูลที่ใช้ในการเลือกแผนหรือกลยุทธ์ที่เหมาะสมในการสอบถามข้อมูลจากฐานข้อมูล โดยระบบจัดการฐานข้อมูลจะมีตัวประมวลผลที่เรียกว่า ตัวประมวลผลสอบถาม (query processor) ทำหน้าที่ในการเลือกแผนหรือกลยุทธ์ที่เหมาะสมเพื่อใช้ในการประมวลผล

ตอนที่ 8.2 การทำกระบวนการสอบถามข้อมูล และการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

หัวเรื่อง

เรื่องที่ 8.2.1 ขั้นตอนของการทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

แนวคิด

1. กระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุดมีขั้นตอนซึ่งประกอบด้วย การตรวจสอบไวยากรณ์และความถูกต้องของคำสั่งเอสคิวแอลตามหลักภาษา การเลือกแผนการทำงานที่จะให้ได้คำตอบที่ดีที่สุด การสร้างโค้ดที่จะทำงานกับแผนที่ถูกเลือก และขั้นตอนสุดท้ายคือการให้โค้ดของแบบสอบถามที่สร้างไว้ให้ทำงานเพื่อจะได้ข้อมูลตามที่ผู้ใช้ต้องการจากการป้อนภาษาเอสคิวแอลเข้าไป
2. เพื่อที่จะให้การเลือกแผนการทำงานที่ดีที่สุด ต้องอาศัยเทคนิคหรือกฎเกณฑ์บางอย่าง เทคนิคนี้จะช่วยประหยัดเวลาให้การเลือกแผนการทำงาน และได้แผนการทำงานที่ดีที่สุด

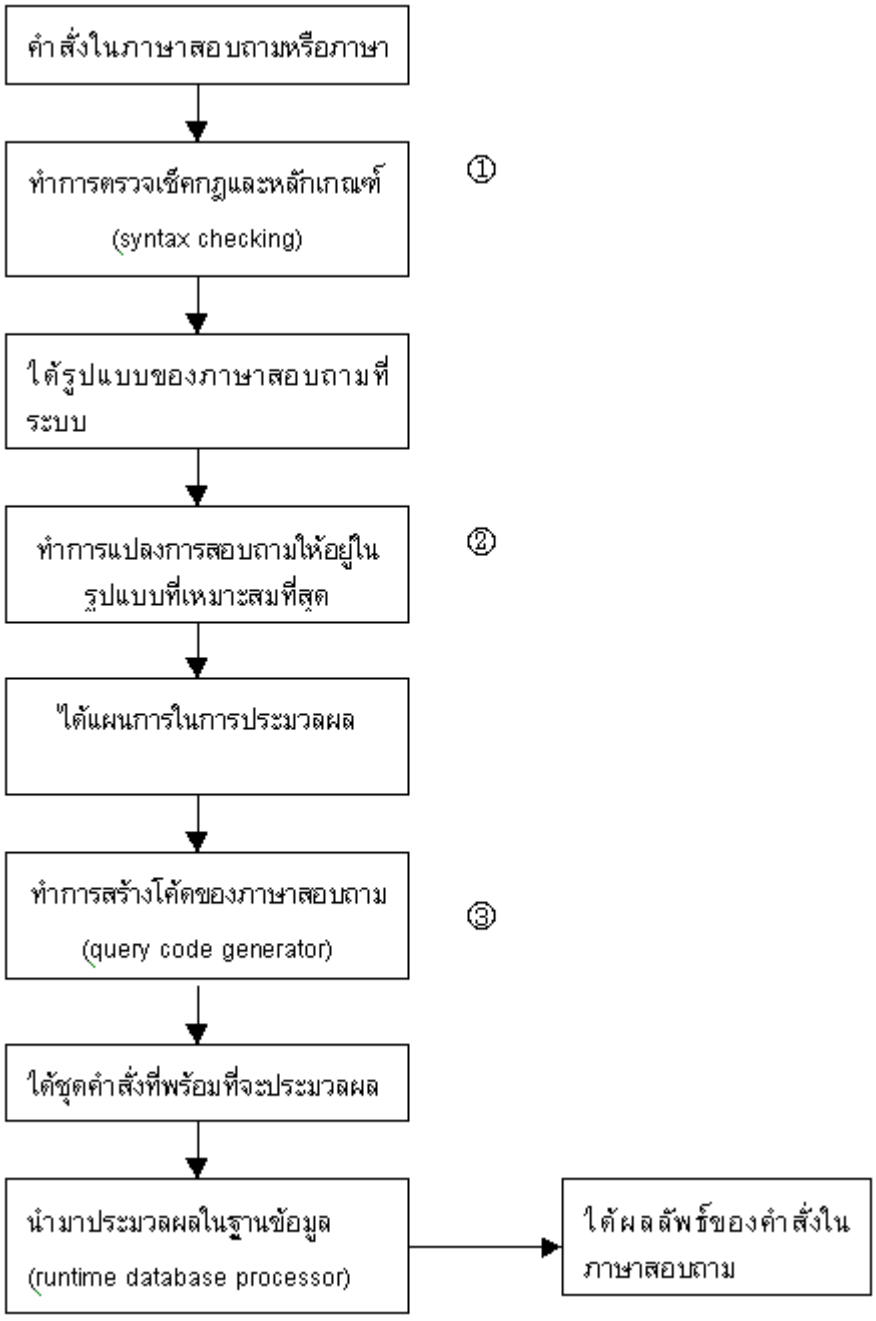
วัตถุประสงค์

หลังจากศึกษาตอนที่ 8.2 แล้ว นักศึกษาสามารถ

1. อธิบายขั้นตอนของกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด
2. บอกเทคนิคและนำเสนอเทคนิคที่สามารถนำมาประยุกต์ใช้ในการเลือกแผนการทำงาน
3. บอกความหมายของสัญลักษณ์ที่ใช้อยู่ในรูปแบบของโมเดลทางคณิตศาสตร์ ซึ่งได้จากภาษาเอสคิวแอลถูกแปลงให้อยู่ในรูปของภาษาระดับต่ำ

เรื่องที่ 8.2.1 ขั้นตอนของการทำกระบวนการสอบถามข้อมูล และการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

การประมวลผลสอบถามจะทำการประมวลผลเป็นขั้นตอนต่าง ๆ โดยจะนำคำสั่งภาษาสอบถามที่ผู้ใช้ต้องการมาตรวจสอบกับกฎและไวยากรณ์ (rule of grammars) ของภาษาสอบถามนั้น ๆ ระบบการจัดการฐานข้อมูลจะใช้กลยุทธ์การประมวลผล (execution strategy) สำหรับหาผลลัพธ์ของคำสั่งภาษาสอบถามนั้น



ๆ จะทำให้ได้ผลลัพธ์ในเวลาอันรวดเร็ว ขั้นตอนในการประมวลผลสอบถามดังภาพที่ 8.2.1 มีขั้นตอนดังนี้

ภาพที่ 8.2.1 ขั้นตอนการประมวลผลในภาษาสอบถาม

ขั้นตอนที่ 1 การตรวจเช็คกฎไวยากรณ์ของภาษา (syntax checking) เมื่อผู้ใช้ได้ทำการป้อนคำสั่งภาษาสอบถาม ระบบจัดการฐานข้อมูลโดย ตัวประมวลผลสอบถาม (query processor) จะทำการตรวจเช็คกฎและหลักเกณฑ์ (syntax checking) ของภาษาสอบถามนั้น เมื่อตรวจเช็คแล้วหากพบว่าคำสั่งภาษาสอบถามที่ป้อนเข้ามายังระบบมีความผิดพลาดไม่เป็นไปตามกฎเกณฑ์ของภาษาสอบถาม ระบบจัดการฐานข้อมูลโดยตัวประมวลผลสอบถามก็จะส่งข้อความไปยังผู้ใช้ให้ผู้ใช้ปรับแก้ให้ถูกต้อง แต่ถ้าตรวจเช็คแล้วคำสั่งนั้นถูกต้องตามกฎไวยากรณ์ของภาษาสอบถาม จะแปลงทำให้ได้รูปแบบของภาษาสอบถาม (intermediate from of query) เพื่อให้ระบบจัดการฐานข้อมูลสามารถเข้าใจได้ ซึ่งผลที่แปลงจะอยู่ในรูปของสมการพีชคณิตแบบสัมพันธ์ (relational algebraic expression) เพื่อให้ ระบบจัดการฐานข้อมูลเข้าใจและทำตามผลที่ได้นั้น ตัวอย่างเช่น ต้องการหารายชื่อนักศึกษาที่ลงทะเบียนในวิชาฐานข้อมูล ซึ่งสามารถเขียนเป็นภาษา เอสคิวแอลได้ดังนี้

```
select Std_Name, Course_Name
from  STUDENT, REGISTRATION, COURSE
where STUDENT.Std_ID  = REGISTRATION.Std_ID and
          COURSE. = REGISTRATION.Course# and
          REGISTRATION.Course# = 'Database';
```

เมื่อป้อนคำสั่งดังกล่าวข้างตอนเข้าสู่ระบบ ระบบจะทำการตรวจเช็คโดยดูว่าคำสั่งที่ผู้ใช้ป้อนเข้าสู่ระบบนั้นมีความผิดพลาดหรือไม่ ถ้าพบว่าผิดพลาดระบบจะส่งข้อความแจ้งให้ผู้ใช้ทราบเพื่อปรับแก้ให้ถูกต้อง แต่ถ้าคำสั่งที่ป้อนเข้าสู่ระบบนั้นถูกต้องแล้ว ระบบจะสร้างสมการพีชคณิตแบบสมพันธ์ดังนี้

$$\pi_{Std_Name}(\sigma_{Course_Name = Database}(STUDENT \bowtie REGISTRATION \bowtie COURSE))$$

ขั้นตอนที่ 2 การแปลงการสอบถามให้อยู่ในรูปที่เหมาะสมที่สุด (query optimizer) เมื่อได้สมการพีชคณิตแบบสัมพันธ์แล้วระบบการจัดการฐานข้อมูลจะทำการประมวลผลสมการรพีชคณิตแบบสัมพันธ์นั้น เพื่อให้ได้แผนการเข้าถึงข้อมูลในการประมวลผล การที่จะให้ได้แผนการเข้าถึงข้อมูล (access) นั้นจะมีวิธีการเข้าถึงฐานข้อมูลได้หลายวิธี ระบบการจัดการฐานข้อมูลจะทำการหาวิธีที่เหมาะสมในการเข้าถึงฐานข้อมูลโดยจะมีเทคนิคต่าง ๆในการเข้าถึงข้อมูลที่ใช้เวลาน้อยที่สุด หรือเรียกว่าการทำออพติไมเซชั่น (optimisation) ซึ่งเทคนิคในการเข้าถึงข้อมูลที่เหมาะสมเหล่านี้ เช่น เทคนิคการซอร์ตและการเมอร์จ (sorting and merge method) เป็นการเรียงลำดับข้อมูลในตารางข้อมูลโดยในระบบจัดการฐานข้อมูลจะมีโปรแกรมสำหรับจัดเรียงลำดับข้อมูล เพื่อให้การประมวลผลประหยัดเวลาในการประมวลผลลง เทคนิคดีคอมโพสิชั่น (decomposition method) เป็นการแตกคำสั่งในภาษาสอบถามให้เป็นโมดูลย่อย ๆเพื่อง่ายต่อการประมวลผล หรือเทคนิคโอเปอร์เรเตอร์กราฟ (operator graph) เป็นการนำสมการพีชคณิตที่ได้จากการตรวจเช็คกฎไวยากรณ์ของภาษา แล้วนำมาหาทางการประมวลผลที่ทำให้การประมวลผลนั้นใช้เวลาน้อย เทคนิคต่าง ๆ เหล่านี้จะทำให้ได้แผนการในการประมวลผลที่เหมาะสมว่าระบบจะต้องประมวลผลอย่างไร เพื่อให้ได้วิธีการเข้าถึงฐานข้อมูลที่มีประสิทธิภาพและใช้เวลาในการประมวลผลน้อยที่สุด

ขั้นตอนที่ 3 การสร้างชุดคำสั่งของภาษาสอบถาม (query code generator) เมื่อได้วิธีที่เหมาะสมที่สุดในการเข้าถึงฐานข้อมูลแล้ว ระบบจัดการฐานข้อมูลโดยตัวประมวลผลสอบถามจะทำการสร้างชุดคำสั่งของภาษาสอบถาม ทำให้ได้ชุดคำสั่ง (source code) ที่พร้อมจะประมวลผล

ขั้นตอนที่ 4 การประมวลผลในฐานข้อมูล (runtime database processor) เมื่อได้ชุดคำสั่งที่พร้อมจะประมวลผลแล้ว ตัวประมวลผลสอบถามก็จะทำการคอมไพล์ (compile) หรือแปลคำสั่ง (interpret) และประมวลผล (executed) ชุดคำสั่งจนได้ผลลัพธ์

กิจกรรม 8.2

1. ขั้นตอนของการประมวลผลสอบถามประกอบด้วยอะไรบ้าง

แนวตอบกิจกรรม 8.2

1. ขั้นตอนของการประมวลผลสอบถามประกอบด้วย
    1. การตรวจเช็คกฎไวยากรณ์และหลักเกณฑ์ของภาษา
    2. การแปลงการสอบถามให้อยู่ในรูปที่เหมาะสมที่สุด
    3. การสร้างชุดคำสั่งของภาษาสอบถาม
    4. การประมวลผลในฐานข้อมูล

ตอนที่ 8.3 การประเมินค่าใช้จ่ายที่เกิดขึ้นของการทำกระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

หัวเรื่อง

เรื่องที่ 8.3.1 การกำหนดฟังก์ชันของค่าใช้จ่าย

เรื่องที่ 8.3.2 การเปรียบเทียบเทคนิคที่นำมาประยุกต์ใช้ที่ก่อให้เกิดค่าใช้จ่ายที่ต่ำที่สุดในการทำที่ก่อให้เกิดค่าใช้จ่ายที่ต่ำที่สุดและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

แนวคิด

1. กระบวนการสอบถามข้อมูลและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุดนั้น ขั้นตอนในการเลือกแผนการทำงานจะมีค่าใช้จ่ายเกิดขึ้น ค่าใช้จ่ายนั้นคือการดึงข้อมูลจากดิสค์, การดึงข้อมูลจากดิสค์เข้ามาเก็บในหน่วยความจำ ดังนั้นเทคนิคที่นำมาประยุกต์ใช้ในการเลือกแผนการทำงานที่ดีที่สุดคือ เทคนิคที่เลือกมาแผนการทำงานนั้นที่ก่อให้เกิดค่าใช้จ่ายที่น้อยที่สุด ความเป็นไปได้ที่แผนการทำงานจะมีหลายแผนการ และแต่ละแผนจะก่อให้เกิดค่าใช้จ่ายที่มากน้อยต่างกันจึงต้องมีการเปรียบเทียบค่าใช้จ่ายที่เกิดขึ้น

วัตถุประสงค์

เมื่อศึกษาบทเรียนที่จบแล้ว นักศึกษาสามารถ

1. อธิบายความหมายของฟังก์ชั่นของค่าใช้จ่ายของแผนการทำงานในกระบวนการสอบถามข้อมูล
2. ทราบถึงวิธีการเลือกแผนการทำงานที่ก่อให้เกิดค่าใช้จ่ายที่น้อยที่สุดได้

เรื่องที่ 8.3.1 การกำหนดฟังก์ชันของค่าใช้จ่าย

ค่าใช้จ่ายในการประมวลผลสอบถามจะมีค่าใช้จ่ายที่เกิดขึ้นดังนี้

1. ค่าใช้จ่ายที่เกิดขึ้นจากการเข้าถึงข้อมูลในฐานข้อมูลที่เก็บอยู่ในดิสค์ซึ่งเป็นหน่วยเก็บข้อมูลสำรอง ค่าใช้จ่ายที่เกิดขึ้นจากการเข้าถึงข้อมูลนั้นประกอบด้วย การค้นหา การอ่าน และการเขียนข้อมูลลงในดิสค์ ค่าใช้จ่ายในการค้นหาข้อมูลขึ้นอยู่กับการจัดโครงสร้างของการจัดเก็บแฟ้มข้อมูลเชิงกายภาพในฐานข้อมูลว่าเป็นอย่างไร เช่น แฟ้มข้อมูลแบบเรียงลำดับ, แฟ้มข้อมูลแบบเข้าถึงโดยตรง, โครงสร้างของตัวดัชนีและตัวดัชนีรอง ในฐานข้อมูลขนาดใหญ่ การประมวลผลสอบถามจะต้องพยายามลดค่าใช้จ่ายของการเข้าถึงข้อมูลให้มีค่าใช้จ่ายน้อยที่สุด ซึ่งขึ้นอยู่กับว่าแผนหรือกลยุทธ์ในการประมวลผลอันไหนที่จะทำให้ค่าใช้จ่ายนี้เกิดขึ้นน้อยที่สุด
2. ค่าใช้จ่ายที่เกิดจากแฟ้มข้อมูลชั่วคราว (intermediate file) ที่ต้องเก็บอยู่ในดิสค์ แฟ้มข้อมูลชั่วคราวจะถูกสร้างขึ้นโดยกลยุทธ์การประมวลผลของการประมวลผลสอบถาม
3. ค่าใช้จ่ายที่เกิดจากการคำนวณของการค้นหาข้อมูล การเรียงลำดับของระเบียนข้อมูล การนำข้อมูลมารวมกัน และการคำนวณค่าต่างของข้อมูลที่เป็นตัวเลขในแต่ละเขตข้อมูล(field) ซึ่งเกิดขึ้นในหน่วยความจำหลักในระหว่างการทำการประมวลผลสอบถาม ในฐานข้อมูลขนาดเล็กข้อ

มูลส่วนใหญ่ในฐานข้อมูลจะถูกนำมาประมวลผลในหน่วยความจำทั้งหมด ดังนั้นจึงคำนึงการลดค่าใช้จ่ายในการคำนวณในหน่วยความจำ สำหรับฐานข้อมูลขนาดเล็ก

4. ค่าใช้จ่ายที่เกิดจากการส่งภาษาสอบถาม (query) บนฐานข้อมูลที่อยู่ในที่หนึ่งไปยังฐานข้อมูลที่อยู่อีกที่หนึ่งเพื่อทำการประมวลผลสอบถาม และต้องส่งผลลัพธ์ที่เกิดจากการประมวลผลสอบถามจากฐานข้อมูลนั้นกลับไปยังฐานข้อมูลที่ภาษาสอบถามนั้นส่งมา ในฐานข้อมูลแบบกระจาย (distributed databases) เมื่อฐานข้อมูลแต่ละฐานข้อมูลตั้งอยู่กระจายในหลายๆที่ที่ห่างกันออกไป การส่งข้อมูลไปประมวลผลผ่านระบบเครือข่ายจึงต้องมีค่าใช้จ่ายเกิดขึ้น ดังนั้นในระบบฐานข้อมูลแบบกระจาย การประมวลผลสอบถามจึงต้องคำนึงถึงค่าใช้จ่ายในส่วนนี้ให้เกิดน้อยที่สุด

เรื่องที่ 8.3.2 การเปรียบเทียบเทคนิคที่นำมาประยุกต์ใช้ที่ก่อให้เกิดค่าใช้จ่ายที่ต่ำที่สุดในการทำที่ก่อให้เกิดค่าใช้จ่ายที่ต่ำที่สุดและการสอบถามข้อมูลที่ได้คำตอบที่ดีที่สุด

สมมุติให้ระบบฐานข้อมูลของมหาวิทยาลัยแห่งหนึ่งประกอบด้วยตารางข้อมูล 4 ตาราง และแอตทริบิวต์ต่าง ๆ ดังนี้

- ตารางนักศึกษา STUDENT (Std #, Std_Name)
- ตารางลงทะเบียนREGISTRATION (Std#, Course#)
- ตารางเกรด GRADE (Std#, Course#, Grade)
- ตารางวิชา COURSE (Course#, Course_Name, Instructor)

โดยในแต่ละตารางจะมีจำนวนข้อมูลดังนี้

- ตาราง STUDENT มีข้อมูลนักศึกษาจำนวน 40,000 แถว
- ตาราง REGISTRATION มีข้อมูล ตามจำนวนที่นักศึกษาลงทะเบียนเรียน แต่ไม่ใช่นักศึกษาทั้งหมด ในตัวอย่างนี้จะสมมุติว่า มีวิชาที่นักศึกษาต้องลงทะเบียนทั้งหมด 10 วิชา ต่อปี จะมีจำนวนข้อมูลทั้งหมดประมาณ 400,000 แถว
- ตาราง GRADE ถ้าสมมุติให้มีข้อมูลเฉลี่ย 10 วิชา ต่อนักศึกษา 1 คน จะมีข้อมูลทั้งหมด 400,000 แถว
- ตาราง COURSE มีข้อมูลวิชาบรรจุอยู่ 5,000 วิชา

ถ้าต้องการหารายชื่อนักศึกษาที่ลงทะเบียนในวิชา (Couse#) ที่เรียงลำดับอยู่หหลังวิชา COMP300 จากโจทย์ข้างต้นในภาษา SQL ผู้ใช้สามารถเลือกใช้คำสั่งได้หลายคำสั่งดังนี้คือ

คำสั่งที่ 1

```
select Std_Name, Course_Name
from STUDENT, REGISTRATION, COURSE
where STUDENT.Std# = REGISTRATION.Std# and
    COURSE.Course# = REGISTRATION.Course# and
```

```
        REGISTRATION.Course# > COMP300
```

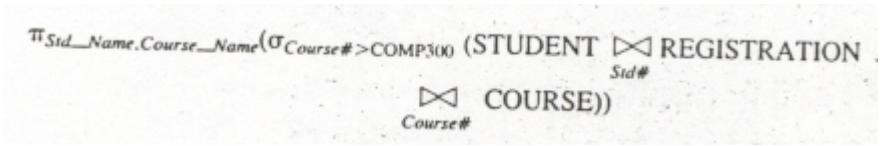
คำสั่งที่ 2

```
select Std_Name, cl.Course_Name
from STUDENT, REGISTRATION, COURSE c1
where STUDENT.Std# = REGISTRATION.Std# and
    REGISTRATION.course# in
    (select c2.Course#
    from COURSE c2
    where c2.Course# > COMP300 and
    cl.Course# = c2.Course#)
```

คำสั่งที่ 3

```
select Std_Name, cl.Course_Name
from STUDENT, COURSE c1
where STUDENT.Std# in
    (Select REGISTRATION.Std#
    from REGISTRATION
    where REGISTRATION.course# in
    (select c2.Course#
    from COURSE c2
    where c2.Course# > COMP300 and
    cl.Course# = c2.Course#))
```

จากภาษา SQL ทั้ง 3 คำสั่งนี้เมื่อผ่านส่วนตรวจสอบภาษา (parser) ที่อยู่ในตัวประมวลผลสอบถามแล้วจะได้รูปแบบของพีชคณิตแบบสัมพันธ์ (relational algebra) ดังนี้



คำสั่งที่ 1

คำสั่งที่ 2

$$\pi_{Std_Name, Course_Name}(\text{STUDENT} \underset{Std\#}{\bowtie} (\sigma_{Course\# > COMP300}(\text{REGISTRATION} \underset{Course\#}{\bowtie} \text{COURSE}))$$

คำสั่งที่ 3

$$\pi_{Std_Name, Course_Name}(\text{STUDENT} \underset{Std\#}{\bowtie} (\sigma_{Course\# > COMP300}\ \text{REGISTRATION}) \underset{Course\#}{\bowtie} (\sigma_{Course\# > COMP300}\ \text{COURSE})$$

จากคำสั่งทั้ง 3 คำสั่งนี้จะทำการเปรียบเทียบว่าคำสั่งใดที่ทำการประมวลผลโดยการดึงข้อมูลจากฐานข้อมูลและขนาดของข้อมูลที่ได้ในแต่ละขั้นตอนในการประมวลผลน้อยที่สุด

จากคำสั่งที่ 1 จะมีการดึงข้อมูลและการประมวลผลข้อมูลได้ตามตารางที่ 8.3.2.1 มีขั้นตอนดังนี้

ตารางที่ 8.3.2.1 ข้อมูลการประมวลผลในแต่ละขั้นตอนของคำสั่งที่ 1

$$\pi_{Std_Name, Course_Name}(\sigma_{Course\# > COMP300}(\text{STUDENT} \underset{Std\#}{\bowtie} \text{REGISTRATION} \underset{Course\#}{\bowtie} \text{COURSE}))$$

Operation	Processing cost if relations		Estimated size of result
	Not sorted	sorted	
Join of STUDENT and REGISTRATION	40,000 * 400,000	40,000 + 400,000	400,000 tuples
Join of this result with COURSE	5,000 * 400,000	5,000 + 400,000	400,000 tuples
Selection from result of *Course#* > COMP300	400,000	400,000	40,000 tuples
Projection on *Std_Name*, *Course_Name*	40,000	40,000	40,000 tuples

ขั้นที่ 1 เป็นการจอยน์ (join) ตาราง STUDENT กับตาราง REGISTRATION ซึ่งในตาราง STUDENT จะมีข้อมูลอยู่ 40,000 แถว และตาราง REGISTRATION มีข้อมูล 400,000 แถว เมื่อจอยน์กันแล้วจะทำให้ได้ผลลัพธ์ของการประมวลผลออกมาประมาณ 400,000 แถว ครั้ง ถ้าในกรณีที่ข้อมูลนั้นไม่ได้จัดเรียง (not sorted) ผลลัพธ์ที่ได้นี้จะต้องประมวลผลถึง 40,000 *[1] 400,000 ครั้ง แต่ถ้าข้อมูลนั้นถูกจัดเรียง (sorted) ก็จะประมวลผลเพียง 40,000 + 400,000 ครั้ง

ขั้นที่ 2 เป็นการจอยน์ (join) ระหว่างผลลัพธ์ที่ได้จากขั้นที่ 1 กับจำนวนแถวที่อยู่ในตาราง COURSE ซึ่งในกรณีที่ไม่มีการจัดเรียงข้อมูลจะทำให้ต้องประมวลผลถึง 5,000 * 400,000 ครั้ง ถ้ามีการจัดเรียงข้อมูลจะประมวลผลเพียง 5,000 + 400,000ครั้ง ซึ่งจะได้ผลลัพธ์ในขั้นตอนนี้ประมาณ 400,000 แถว

ขั้นตอนที่ 3 จากผลลัพธ์ที่ได้ในขั้นที่ 2 จะทำการเลือกวิชาที่มีลำดับมากกว่า COMP300 โดยในการเลือกนี้จะ ถ้าเราสมมุติว่าวิชาที่มากกว่า COMP300 จะทำการประมวลผลโดยการเปรียบเทียบเป็นจำนวน 400,000 ครั้ง ซึ่งจะได้ผลลัพธ์ของการประมวลผลในขั้นตอนนี้ประมาณ 40,000 แถว

ขั้นที่ 4 เป็นการจัดกลุ่ม (projection) ของ Std_Name และ Course_Name ซึ่งจะทำการประมวลผลทั้งหมด 40,000 ครั้งและได้ผลลัพธ์ออกมา 40,000 แถว

จากคำสั่งที่ 2 เมื่อพิจารณาการดึงข้อมูลและการประมวลผลข้อมูลตามตารางที่ 8.3.2.2 ทีมีการประมวลผลดังนี้

ตารางที่ 8.3.2.2 ข้อมูลการประมวลผลในแต่ละขั้นตอนของคำสั่งที่ 2

$$\pi_{Std_Name,\ Course_Name}(STUDENT \bowtie_{Std\#} (\sigma_{Course\# > COMP300}(REGISTRATION \bowtie_{Course\#} COURSE)))$$

Operation	Processing cost if relations		Estimated size of result
	Not sorted	sorted	
Join of REGISTRATION and COURSE	5,000 * 400,000	5,000 + 400,000	400,000 tuples
Selection from result of *Course#* > COMP300	400,000	400,000	40,000 tuples
Join of STUDENT and result above	40,000 * 40,000	40,000 + 40,000	40,000 tuples
Projection on *Std_Name, Course_Name*	40,000	40,000	40,000 tuples

ขั้นที่ 1 เป็นการจอยน์ระหว่างตาราง REGISTRATION และตาราง COURSE โดยในกรณีที่ข้อมูลไม่ได้จัดเรียงจะต้องประมวลผลทั้งหมด 5,000 * 400,000 ครั้ง ถ้าข้อมูลจัดเรียงลำดับจะประมวลผลทั้งหมด 5,000 + 400,000 ครั้ง ผลลัพธ์ที่ได้จากขั้นที่ 1 นี้ จะมีประมาณ 400,000 แถว

ขั้นที่ 2 จากผลลัพธ์ที่ได้ในขั้นที่ 1 จะทำการเลือกวิชาที่มากกว่า COMP300 ซึ่งจะต้องประมวลผลถึง 400,000 ครั้ง จะได้ผลลัพธ์ออกมาประมาณ 40,000 แถว

ขั้นที่ 3 เป็นการจอยน์ภายในตาราง STUDENT กับผลลัพธ์ที่ได้ในขั้นที่ 2 เพื่อหา Std# โดยในกรณีที่ข้อมูลไม่ได้จัดเรียงจะประมวลผลทั้งหมด 40,000 * 40,000ครั้ง และจะประมวลผลทั้งหมด 40,000 + 40,000 ครั้ง ในกรณีที่ข้อมูลจัดเรียง ซึ่งจะได้ผลลัพธ์ออกมาประมาณ 40,000 แถว

ขั้นที่ 4 เป็นการจัดกลุ่ม (projection) ของ Std_Name และ Course_Name จากผลลัพธ์ที่ได้ในขั้นที่ 2 ซึ่งมีการประมวลทั้งหมด 40,000 ครั้ง และได้ผลลัพธ์ของการประมวลผลประมาณ 40,000 แถว

จากคำสั่งที่ 3 จะเห็นว่าจะใช้คำสั่งในการเลือก (selection) ก่อนที่จะจอยน์ (joins) ซึ่งเมื่อพิจารณาการดึงข้อมูลและการประมวลผลข้อมูลจะได้ ดังตารางที่ 8.3.2.3

ตารางที่ 8.3.2.3 แสดงข้อมูลการประมวลผลในแต่ละขั้นตอนของคำสั่งที่ 3

$$\pi_{Std_Name,Course_Name}(\text{STUDENT} \underset{Std\#}{\bowtie} (\sigma_{Course\#>COMP300}(\text{REGISTRATION}) \underset{Course\#}{\bowtie} (\sigma_{Course\#>COMP300}\text{COURSE})$$

Operation	Processing cost if relations		Estimated size of result
	Not sorted	sorted	
Selection from COURSE *Course#* > COMP300	5,000	5,000	500 tuples
Selection from REGISTRATION *Course#* > COMP300	400,000	400,000	40,000 tuples
Join of selected tuples from COURSE and REGISTRATION with result above	500 * 40,000	500 + 40,000	40,000 tuples
Join of STUDENT with result above	40,000 * 40,000	40,000 + 40,000	40,000 tuples
Projection on *Std_Name, Course_Name*	40,000	40,000	40,000 tuples

ขั้นที่ 1 เลือก(selection)จากตาราง COURSE โดยรหัสวิชา (Course#) ที่มากกว่า COMP300 ผลของคำสั่งนี้จะทำให้มีการประมวลผล 5,000 ครั้ง และจะได้ผลลัพธ์ออกมาประมาณ 500 แถว

ขั้นที่ 2 เลือกจากตาราง REGITRATION โดยรหัสวิชา (Course#) ที่มากกว่า COMP300 ผลของคำสั่งนี้จะทำให้มีการประมวลผลทั้งหมด 400,000 ครั้ง ได้ผลลัพธ์ประมาณ 400,000 แถว

ขั้นที่ 3 เป็นการจอยน์ระหว่างผลลัพธ์ที่ได้ในขั้นที่ 1 กับผลลัพธ์ที่ได้จากการเลือกรหัสวิชาของขั้นที่ 2 จะทำให้ต้องประมวลผลถึง 500 * 40,000 ครั้งในกรณีที่ไม่ได้เรียงข้อมูลและจำนวน 500 + 40,000 กรณีที่จัดเรียงข้อมูลและจะได้ผลลัพธ์ออกมาประมาณ 40,000 แถว

ขั้นที่ 4 ผลลัพธ์ที่ได้จากขั้นที่ 3 จะจอยน์กับตาราง STUDENT ในกรณีที่ไม่ได้จัดเรียงข้อมูล จะทำต้องมีการประมวลผลทั้งหมด 40,000 * 40,000 ครั้งหรือในกรณีที่จัดเรียงข้อมูลจะประมวลผลทั้งหมด 40,000 + 40,000 ครั้งจะได้ผลลัพธ์ประมาณ 40,000 แถว

ขั้นที่ 5 ผลลัพธ์ที่ได้จากขั้น 4 จะจัดกลุ่มของ Std_Name และ Course_Name ซึ่งจะต้องประมวลผลทั้งหมด 40,000 ครั้ง จะได้ผลลัพธ์ออกมาประมาณ 40,000 ครั้ง

สรุปว่าจากคำสั่งทั้ง 3 คำสั่ง ผู้ใช้ควรเลือกคำสั่งในการประมวลผลคือคำสั่งที่ 3 เพราะจะมีจำนวนของการประมวลผลและผลลัพธ์ในแต่ละขั้นตอนน้อยกว่าคำสั่งอื่น ๆ การเลือกคำสั่งที่ 3 เข้าประมวลผลถือเป็นการทำออพติไมเซชั่น (optimisation) ที่ผู้ใช้เป็นผู้กระทำโดยการเลือกคำสั่งเข้ามาประมวลผล แต่ในระบบการจัดการฐานข้อมูลนั้นเมื่อผู้ใช้ป้อนคำสั่งภาษาสอบถามเข้าไปแล้ว ระบบจะนำคำสั่งนั้นไปประมวลผลซึ่งในขั้นตอนของการประมวลผลนั้นจะมีการการทำออพติไมเซชั่น (optimisation) ระบบจัดการฐานข้อมูลใช้เทคนิคต่าง ๆ ในการประมวลผลโดยคำนึงถึงการดึงข้อมูลมาประมวลผลให้น้อยที่สุดและใช้เวลาน้อยในการประมวลผลน้อยที่สุด

จากตัวอย่างที่กล่าวมาทั้งหมดจะเห็นว่าการหาแผนการเข้าถึงข้อมูล (access plan) หรือการทำออพติไมเซชั่น (optimisation) สามารถทำได้จากผู้ใช้โดยตรงก่อน โดยผู้ใช้จะเป็นเหมือนออฟติไมเซอร์ได้โดยเลือกคำสั่งที่มีการประมวลผลข้อมูลน้อย เพื่อประหยัดเวลาในการประมวลผล ต่อจากนั้นระบบจัดการฐานข้อมูลจะนำคำสั่งที่ผู้ใช้เลือกแล้วมาทำออพติไมเซชั่น (optimisation) ซันอีกครั้งโดยใช้เทคนิคและวิธีการต่าง ๆ ที่ระบบการจัดการฐานข้อมูลนั้น ๆ มีอยู่ เช่น เทคนิคการซอร์ตและการเมอร์จ (sorting and merge method), เทคนิคดีคอมโพสิชั่น (decomposition method) และเทคนิคโอเปอร์เรเตอร์กราฟ (operator graph) ซึ่งเทคนิคต่างๆ เหล่านี้จะทำให้ใช้เวลาในการประมวลผลน้อย

กิจกรรม 8.3

1. ในกระบวนการสอบถามข้อมูลเพื่อให้ได้คำตอบที่ดีที่สุดนั้นจะต้องคำนึงถึงค่าใช้จ่ายอะไรบ้าง

แนวตอบกิจกรรม 8.3

1. จะต้องคำนึงถึงค่าใช้จ่าย

1. ค่าใช้จ่ายที่เกิดขึ้นจากการเข้าถึงข้อมูลในฐานข้อมูลที่เก็บอยู่ในดิสค์ซึ่งเป็นหน่วยเก็บข้อมูลสำรอง

2. ค่าใช้จ่ายที่เกิดจากแฟ้มข้อมูลชั่วคราว (intermediate file) ที่ต้องเก็บอยู่ในดิสค์

3. ค่าใช้จ่ายที่เกิดจากการคำนวณของการค้นหาข้อมูล การเรียงลำดับของระเบียนข้อมูล การนำข้อมูลมารวมกัน และการคำนวณค่าต่างของข้อมูลที่เป็นตัวเลขในแต่ละเขตข้อมูล(field)

4. ค่าใช้จ่ายที่เกิดจากการส่งภาษาสอบถาม (query) บนฐานข้อมูลที่อยู่ในที่หนึ่งไปยังฐานข้อมูลที่อยู่อีกที่หนึ่งเพื่อทำการประมวลผลสอบถาม และต้องส่งผลลัพธ์ที่เกิดจากการประมวลผลสอบถามจากฐานข้อมูลนั้นกลับไปยังฐานข้อมูลที่ภาษาสอบถามนั้นส่งมา