บทที่ 1

บทนำ

1.1 ความเป็นมา

คลังข้อมูลเป็นที่เก็บข้อมูลที่มีขนาดใหญ่จากหลากหลายแหล่งที่มา ซึ่งมีสมบัติ subject-oriented, integrated, time-variant และ nonvolatile [1,2] โดยการเข้าถึงคลังข้อมูลเป็นแบบ Online Analytical Processing (OLAP) [12] ใช้เพื่อสนับสนุนการตัดสินใจของผู้บริหาร การสอบถามบนคลังข้อมูลมักมีความซับซ้อนและเป็นแบบทันทีทันใด (ad hoc) [25] คือไม่ทราบล่วงหน้าว่าผู้ใช้จะถามอะไรบ้าง เช่น ต้องการทราบยอดขายสินค้า A ยี่ห้อ B ให้แก่ลูกค้าที่เป็นสมาชิกระดับที่ 1 ของสาขาที่หนึ่งและสาขาที่สอง ในช่วงเดือนมกราคมถึงมีนาคม ปี 2548 ซึ่งในการตอบคำถามจะต้องทำการค้นหาข้อมูลเป็นจำนวนมาก จึงมีการประมวลผลที่ยาวนานหลายชั่วโมงหรือหลายวัน แต่ในระบบการสนับสนุนการตัดสินใจนั้น การสอบถามจะต้องให้คำตอบที่มีความถูกต้องแม่นยำและรวดเร็ว การเพิ่มความเร็วในการค้นหาข้อมูลสามารถทำได้หลายวิธี เช่น การสร้าง summary table การประมวลผลแบบคู่ขนาน และการทำดัชนี [11,14, 18] สำหรับการสร้าง summary table นั้น จะมีประสิทธิภาพดีก็ต่อเมื่อเป็นการสร้างเตรียมไว้เพื่อตอบคำถามที่มีการกำหนดไว้ก่อน แต่เมื่อมีคำถามที่เรามิได้กำหนดไว้เกิดขึ้น ระบบจะต้องมีการเข้าถึงข้อมูลจริงทำให้ต้องใช้เวลานาน นอกจากนี้เมื่อ base table มีการเปลี่ยนแปลง เราจะต้องทำการสร้าง summary table ใหม่ด้วย ดังนั้นในการสร้าง summary table จึงควรเลือกสร้างไว้สำหรับการสอบถามที่เกิดขึ้นบ่อย และการสร้างแต่ละครั้งต้องใช้พื้นที่และเวลามาก ยิ่งกว่านั้นเราไม่สามารถสร้าง summary table ทั้งหมดได้ การเลือกว่าจะสร้าง summary table ใดจึงเป็นงานที่ยาก[11] สำหรับการประมวลผลแบบคู่ขนานจะช่วยทำให้การประมวลผลเร็วขึ้น แต่จะต้องเสียค่าใช้จ่ายในการเพิ่มฮาร์ดแวร์มากขึ้น ส่วนการทำดัชนีเป็นทางเลือกในการทำให้การประมวลผลมีประสิทธิภาพ โดยมิต้องมีการเสียค่าใช้จ่ายใด ๆ เพิ่มขึ้น การทำดัชนีบนคลังข้อมูลทำให้เราสามารถค้นหาข้อมูลได้เร็วขึ้น เพราะมีการหาตำแหน่งของข้อมูลบนตารางดัชนีก่อนการเข้าถึงข้อมูลจริง [12,14,15,16,17,19,20] ซึ่งการทำดัชนีมีหลายวิธี แต่ละวิธีก็เหมาะสมกับสถานการณ์ที่แตกต่างกัน คือ B-tree เหมาะกับข้อมูลที่มีลักษณะเป็นช่วงค่า เช่น เงินเดือน, แฮชเหมาะกับข้อมูลที่ต้องการการเข้าถึงโดยตรง เช่น รหัสประจำตัว และดัชนีบิตแมปเหมาะกับข้อมูลที่มีคาร์ดินอลิตี้ไม่สูงมาก โดยผู้วิจัยให้ความสนใจการทำดัชนีบิตแมป เนื่องจากดัชนีบิตแมปเป็นเทคนิคที่มีการประมวลผลรวดเร็ว เพราะมีการดำเนินการระดับบิตระหว่างบิตแมปก่อนเข้าถึงข้อมูลจริง และมีรูปแบบการลงรหัสข้อมูลที่ไม่ซับซ้อน

ในงานวิจัยที่ผ่านมา ผู้วิจัยได้ศึกษาเทคนิคการทำดัชนีบิตแมปที่มีอยู่แล้ว เมื่อพิจารณานำมาใช้กับการสอบถามแบบค่าเท่ากันพบว่า เทคนิคที่มีอยู่ยังมีข้อบกพร่องบางประการอยู่ ทั้งในเรื่องของพื้นที่ที่ใช้ในการสร้างดัชนีและเวลาที่ใช้ในการสอบถาม

ในงานวิจัยนี้ ผู้วิจัยจึงได้คิดค้นวิธีในการทำดัชนีบิตแมปแบบใหม่ขึ้น มีชื่อว่า ดัชนีบิตแมปแบบคู่กัน(Dual Bitmap Index) ซึ่งมีประสิทธิภาพในเรื่องของพื้นที่ที่ใช้ในการสร้างดัชนีมากกว่าดัชนีบิตแมปที่เคยมีมา แต่ยังคงมีประสิทธิภาพในเรื่องของเวลาในการตอบคำถามแบบค่าเท่ากัน (เช่น “สินค้า = A” หรือ “ยี่ห้อ = B”) อยู่

1.2 การตรวจเอกสารและงานวิจัยที่เกี่ยวข้อง

1.2.1 การตรวจเอกสาร

การจัดการข้อมูลบนคลังข้อมูลนั้นเป็นสิ่งสำคัญ ต้องมีประสิทธิภาพทั้งในเรื่องของการจัดเก็บและการเข้าถึงข้อมูล โดยเฉพาะอย่างยิ่งการเข้าถึงข้อมูลจะต้องเป็นไปอย่างรวดเร็ว เนื่องจากการสอบถามมักเป็นแบบทันทีทันใด คือไม่ทราบล่วงหน้าว่าผู้ใช้จะสอบถามอะไร วิธีหนึ่งที่นิยมใช้ในการเพิ่มประสิทธิภาพ คือ การทำดัชนีบิตแมป เนื่องจากไม่ต้องเสียค่าใช้จ่ายใดๆ เพิ่ม เพียงแต่จัดการวิธีลงรหัสข้อมูลให้อยู่บนพื้นฐานของบิตแมป เมื่อต้องการเข้าถึงข้อมูล สามารถทำได้ด้วยการเข้าถึงข้อมูลบางส่วนและดำเนินการตรรกะก็เพียงพอ โดยไม่ต้องเข้าถึงข้อมูลทั้งหมด ในงานวิจัยที่ผ่านมามีการคิดค้นเทคนิคการทำดัชนีบิตแมปขึ้นมาหลายวิธี โดยมีการพัฒนาประสิทธิภาพทั้งในด้านการใช้พื้นที่และเวลาให้ดีขึ้นเรื่อยๆ

1.2.2 งานวิจัยที่เกี่ยวข้อง

An Overview of Data Warehousing and OLAP Technology

งานวิจัยนี้ [12] กล่าวถึงความสำคัญของเทคโนโลยีคลังข้อมูล ความแตกต่างระหว่างฐานข้อมูลดำเนินการและคลังข้อมูล สถาปัตยกรรมของเทคโนโลยีคลังข้อมูล การออกแบบเชิงแนวคิด เครื่องมือในการสร้างคลังข้อมูล เครื่องมือในการสอบถามและวิเคราะห์ข้อมูล วิธีการออกแบบคลังข้อมูล การเพิ่มประสิทธิภาพในการสอบถาม การจัดการคลังข้อมูล งานวิจัยในสาขานี้

An Efficient Bitmap Encoding Scheme for Selection Queries

งานวิจัยนี้ [17] กล่าวถึงลักษณะการสอบถามแบบชนิดต่าง ๆ การเพิ่มประสิทธิภาพการสอบถามด้วยการทำดัชนีบิตแมป การทำดัชนีบิตแมปแบบพื้นฐาน (Simple Bitmap Index) ซึ่งมีการสอบถามที่รวดเร็ว มีการนำเสนอเทคนิคการทำดัชนีบิตแมปแบบช่วง

(Interval Bitmap Index) ซึ่งใช้พื้นที่ในการจัดเก็บดัชนีบิตแมปเป็นครึ่งหนึ่งของดัชนีบิตแมปแบบพื้นฐาน แต่ต้องใช้เวลาในการสอบถามมากขึ้น

Encoded Bitmap Indexing for Data Warehouses

งานวิจัยนี้ [16] ได้กล่าวถึงปัญหาที่สำคัญในระบบคลังข้อมูล 3 ประการ คือ การสอบถามมักมีความซับซ้อน ข้อมูลมีจำนวนมหาศาล และมีอัตราการอ่านที่สูงมาก ซึ่งเทคนิคการทำดัชนีบนระบบฐานข้อมูลดำเนินการที่มีอยู่นั้นยังไม่เหมาะสมกับระบบคลังข้อมูล จึงได้มีการนำเสนอเทคนิคการทำดัชนีบิตแมปแบบเข้ารหัส (Encoded Bitmap Index) ซึ่งใช้พื้นที่ในการจัดเก็บดัชนีน้อยกว่าดัชนีบิตแมปชนิดอื่น จึงสามารถใช้ได้กับแอทริบิวต์มีคาร์ดินอลิตี้สูง ๆ แต่ดัชนีบิตแมปแบบเข้ารหัสต้องใช้เวลาในการสอบถามมากกว่าดัชนีบิตแมปชนิดอื่น ๆ

Scatter Bitmap : Space-Time Efficient Bitmap Indexing for Equality and Membership Queries

งานวิจัยนี้ [20] กล่าวถึงประโยชน์ของการทำดัชนีบิตแมปบนคลังข้อมูล การทำดัชนีบิตแมปแบบพื้นฐาน (Simple Bitmap Index) การทำดัชนีบิตแมปแบบช่วง (Interval Bitmap Index) การทำดัชนีบิตแมปแบบเข้ารหัส (Encoded Bitmap Index) มีการนำเสนอเทคนิคการทำดัชนีบิตแมปแบบกระจาย (Scatter Bitmap Index) ซึ่งใช้พื้นที่ในการทำดัชนีบิตแมปน้อยกว่าการทำดัชนีบิตแมปแบบช่วง ในขณะที่ยังคงประสิทธิภาพในการสอบถามแบบค่าเท่ากันและแบบความเป็นสมาชิกอยู่

1.3 วัตถุประสงค์

เพื่อคิดค้นเทคนิควิธีการทำดัชนีบิตแมปแบบใหม่สำหรับการสอบถามแบบค่าเท่ากันจากคลังข้อมูล ที่มีประสิทธิภาพทั้งในด้านการใช้พื้นที่ที่ใช้ในการจัดเก็บดัชนีและเวลาที่ใช้ในการสอบถามดีกว่าเทคนิคที่เคยมีมา

1.4 วิธีการดำเนินการวิจัย

1. ศึกษาแนวคิดที่เกี่ยวข้องกับระบบคลังข้อมูล และดัชนีบิตแมปทั้ง 4 ชนิด ได้แก่ ดัชนีบิตแมปแบบพื้นฐาน ดัชนีบิตแมปแบบช่วง ดัชนีบิตแมปแบบกระจาย และดัชนีบิตแมปแบบเข้ารหัส

2. วิเคราะห์และออกแบบเทคนิคใหม่ในการทำดัชนี สำหรับการสอบถามแบบค่าเท่ากันจากคลังข้อมูล ซึ่งเทคนิคใหม่ที่ได้นี้เรียกว่า ดัชนีบิตแมปแบบคู่กัน

3. กำหนดรูปแบบการประเมินประสิทธิภาพของดัชนีบิตแมปแบบคู่กันที่คิดค้นขึ้นใหม่ เปรียบเทียบกับดัชนีบิตแมปทั้ง 4 ชนิด คือ ดัชนีบิตแมปแบบพื้นฐาน ดัชนีบิตแมป

แบบช่วง ดัชนีบิตแมปแบบกระจาย และดัชนีบิตแมปแบบเข้ารหัส โดยทำการประเมินประสิทธิภาพทั้งในด้านการใช้พื้นที่สำหรับจัดเก็บดัชนีและเวลาที่ใช้ในการสอบถามแบบค่าเท่ากัน (Equality Query) ดังขั้นตอนต่อไปนี้

3.1 ทำการศึกษาข้อมูลสำหรับใช้ในการทดสอบ ซึ่งเป็นข้อมูลมาตรฐานจาก TPC-H Benchmark [25]

3.2 ติดตั้งโปรแกรมสำหรับรันผลการสร้างข้อมูลทดสอบจากการวัดเปรียบเทียบสมรรถนะ TPC-H บนระบบปฏิบัติการ Linux Red Hat 9.0 โดยจัดเก็บฐานข้อมูลทดสอบในรูปของ Flat File

3.3 จัดเตรียมข้อมูลทดสอบ โดยการเลือกเฉพาะแอทริบิวต์ที่จะนำมาทำดัชนี และเปลี่ยนรูปค่าข้อมูลเป็นจำนวนเต็มที่เรียงต่อเนื่องกัน โดยเริ่มจาก 0 (ดูภาคผนวก)

3.4 ออกแบบขั้นตอนวิธีในการสร้างดัชนีบิตแมปทั้ง 5 ชนิด คือ ดัชนีบิตแมปแบบพื้นฐาน ดัชนีบิตแมปแบบช่วง ดัชนีบิตแมปแบบกระจาย ดัชนีบิตแมปแบบเข้ารหัส และดัชนีบิตแมปแบบคู่กัน (ดูภาคผนวก)

3.5 พัฒนาโปรแกรมเพื่อสร้างดัชนีบิตแมปทั้ง 5 ชนิด ตามขั้นตอนที่ออกแบบไว้ตามข้อ 3.4 โดยใช้ตัวแปลภาษาซี

4. ออกแบบขั้นตอนวิธีการสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปทั้ง 5 ชนิด คือ ดัชนีบิตแมปแบบพื้นฐาน ดัชนีบิตแมปแบบช่วง ดัชนีบิตแมปแบบกระจาย ดัชนีบิตแมปแบบเข้ารหัส และดัชนีบิตแมปแบบคู่กัน (ดูภาคผนวก)

5. พัฒนาโปรแกรมเพื่อสอบถามแบบค่าเท่ากันบนดัชนีบิตแมปทั้ง 5 ชนิด ตามขั้นตอนที่ออกแบบไว้ตามข้อ 4 โดยใช้ตัวแปลภาษาซี

6. ประเมินประสิทธิภาพการใช้พื้นที่ในการจัดเก็บดัชนี (Space) ของดัชนีบิตแมปทั้ง 5 ชนิด ด้วยการเขียนกราฟเปรียบเทียบพื้นที่ที่ใช้ในการจัดเก็บดัชนีแต่ละชนิด โดยใช้โปรแกรม MATLAB

7. ประเมินประสิทธิภาพการใช้เวลาในการสอบถามแบบค่าเท่ากัน (Time) บนดัชนีบิตแมปทั้ง 5 ชนิด ด้วยการรันโปรแกรมที่พัฒนาขึ้นตามข้อ 5 แล้วบันทึกเวลาในการสอบถามของดัชนีบิตแมปแต่ละชนิด

8. ประเมินเกี่ยวกับการแลกเปลี่ยนระหว่างประสิทธิภาพของพื้นที่กับเวลา (Space-Time Trade-off) สำหรับการสอบถามแบบค่าเท่ากันบันดัชนีบิตแมปแต่ละชนิด

9. วิเคราะห์และสรุปผลการประเมินประสิทธิภาพดัชนีบิตแมปแบบคู่กันที่คิดค้นขึ้น กับดัชนีเปรียบเทียบทั้ง 4 ชนิด

1.5 ขอบเขตงานวิจัย

1. ศึกษาดัชนีบิตแมปแบบเดิมที่เคยมีอยู่ ได้แก่ ดัชนีบิตแมปแบบพื้นฐาน ดัชนีบิตแมปแบบช่วง ดัชนีบิตแมปแบบกระจาย และดัชนีบิตแมปแบบเข้ารหัส

2. วิเคราะห์และออกแบบเทคนิคเพื่อสร้างดัชนีแบบใหม่สำหรับการค้นหาข้อมูลแบบค่าเท่ากันจากคลังข้อมูล

3. ประเมินประสิทธิภาพของการใช้พื้นที่ในการจัดเก็บดัชนี (Space) การใช้เวลาในการสอบถามแบบค่าเท่ากัน (Time) และการแลกเปลี่ยนระหว่างประสิทธิภาพของพื้นที่กับเวลา (Space-Time Trade-off) บนดัชนีบิตแมปที่สร้างขึ้นใหม่และดัชนีบิตแมปที่เคยมีมา

1.6 ขั้นตอนการดำเนินงาน

1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้องและกำหนดขอบเขตของปัญหาให้ชัดเจน

2. วิเคราะห์และออกแบบเทคนิคการสร้างดัชนีบิตแมปสำหรับการสอบถามแบบค่าเท่ากันจากคลังข้อมูล

3. กำหนดรูปแบบการประเมินประสิทธิภาพของดัชนีที่สร้างขึ้นใหม่ ซึ่งเรียกว่าดัชนีบิตแมปแบบคู่กัน และดัชนีเปรียบเทียบทั้ง 4 ชนิด คือ ดัชนีบิตแมปแบบพื้นฐาน ดัชนีบิตแมปแบบช่วง ดัชนีบิตแมปแบบกระจาย และดัชนีบิตแมปแบบเข้ารหัส

4. ศึกษาและวิเคราะห์หาเครื่องมือสำหรับใช้ประเมินดัชนีที่สร้างขึ้นใหม่และดัชนีเปรียบเทียบ

5. พัฒนาดัชนีการค้นหาข้อมูล ตามที่ได้ทำการออกแบบไว้

6. ดำเนินการประเมินดัชนีที่สร้างขึ้นใหม่กับดัชนีเปรียบเทียบ

7. สรุปและวิเคราะห์ผลการประเมินดัชนีที่สร้างขึ้นใหม่กับดัชนีเปรียบเทียบ

8. จัดทำเอกสารประกอบการวิจัย

1.7 ระยะเวลาดำเนินงานและแผนการดำเนินงาน

ระยะเวลาดำเนินงาน ตุลาคม พ.ศ. 2548 - มีนาคม พ.ศ. 2550

ตาราง 1-1 แสดงระยะเวลาดำเนินงาน

ขั้นตอนที่	เดือน																	
	2548			2549												2550		
	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3
1	←	—	→															
2			←	—	→													
3					←	—	→											
4						←	—	→										
5								←	—	—	→							
6											←	—	→					
7													←	—	→			
8														←	—	—	—	→

1.8 เครื่องมือและอุปกรณ์

ด้านฮาร์ดแวร์

เครื่องคอมพิวเตอร์ สำหรับที่ใช้ในการเตรียมข้อมูลจาก TPC-H และพัฒนาโปรแกรมเพื่อประเมินประสิทธิภาพของเวลาที่ใช้ในการสอบถามแบบค่าเท่ากับบนดัชนีบิตแมปทั้ง 5 ชนิด ซึ่งมีสมรรถนะดังนี้

ซีพียู : Intel(R) Celeron(R)
หน่วยความจำ : 512 MB
ฮาร์ดดิสก์ : 40 GB

ด้านซอฟต์แวร์

- ระบบปฏิบัติการ Linux Red Hat 9.0
- ตัวแปลภาษาซี (C Compiler) สำหรับใช้ในการพัฒนาโปรแกรมเพื่อวัดประสิทธิภาพของเวลาที่ใช้ในการสอบถาม

1.9 ประโยชน์ที่คาดว่าจะได้รับ

ได้วิธีการทำดัชนีบิตแมปแบบใหม่ที่มีการแลกเปลี่ยนระหว่างประสิทธิภาพของพื้นที่กับเวลา (Space-Time Trade-off) อย่างคุ้มค่า สำหรับการสอบถามค่าเท่ากันจากคลังข้อมูล