บทที่ 4

การทดลองและผลการทดลอง

การทดลองในบทนี้มีวัตถุประสงค์เพื่อทำการทดสอบประสิทธิภาพด้วยการวัดความเร็วของขั้นตอนวิธีในการขั้นหาความซ้ำซ้อนของข้อมูลภายในฐานข้อมูล

แนวทางของการแก้ปัญหาในงานวิจัยครั้งนี้ คือการการปรับปรุงประสิทธิภาพในกระบวนการค้นหาความซ้ำซ้อนของข้อมูล โดยใช้วิธีการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอล (PL/SQL) และการใช้วิธีการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมาเอง (UDF) จากบทที่ 3 ซึ่งมีเทคนิคที่ใช้ในการประมวลผล 2 เทคนิค คือการเขียนการสอบถามใหม่ (Query Rewriting) และการสร้างดัชนี (Indexing) เทคนิคดังกล่าว สามารถแบ่งได้ออกเป็น 4 แนวทางในการทดลองเพื่อเปรียบเทียบประสิทธิภาพ ดังต่อไปนี้

1. แนวทางการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมาเอง
2. แนวทางการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมาเอง และมีการใช้เทคนิคการสร้างดัชนีข้อมูลเข้าร่วมด้วย
3. แนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด
4. แนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด มีการใช้เทคนิคการสร้างดัชนีข้อมูล และเทคนิคการเขียนการสอบถามใหม่เข้าร่วมด้วย

ในงานวิจัยนี้จะทำการเปรียบเทียบประสิทธิภาพในการค้นหาความซ้ำซ้อนของข้อมูล จากแนวทางที่ได้เสนอไปทั้ง 4 แนวทาง โดยจะทำการเปรียเทียบจากเวลาที่ใช้ในการค้นหาความซ้ำซ้อน

4.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่จะนำมาใช้ในการทดลองสำหรับค้นหาความซ้ำซ้อนของข้อมูล คือชุดข้อมูลของทีซีพีเอช เบนมาค (TPC Benchmark™H) จากฐานข้อมูลข้อมูลดังกล่าวจะประกอบไปด้วยชุดของข้อมูลที่มีหลายขนาดด้วยกัน ซึ่งงานในงานวิจัยจัยนี้ต้องการที่จะใช้ชุดข้อมูลที่มีขนาดแตกต่างกัน 5 ขนาด เมื่อได้ชุดข้อมูลทั้ง 5 ขนาดแล้วจะต้องนำชุดข้อมูลดังกล่าวไปผ่านกระบวนการต่างๆ ที่ได้อธิบายไว้ในบทที่ 2 ซึ่งเป็นขั้นตอนวิธีในการเตียมข้อมูลให้พร้อมที่จะนำมาใช้ในการทดลอง เพื่อค้นหาความซ้ำซ้อนของข้อมูล

จากชุดของข้อมูลที่ได้มาเมื่อทำการเลือกขนาดของฐานข้อมูลออกมาจำนวน 5 ขนาดแล้ว เพื่อให้เกิดความครอบคลุมกับสถานการณ์ที่อาจเกิดขึ้นจริง เมื่อนำมาใช้งานในการค้นหาความซ้ำซ้อนของข้อมูลจึงได่แบ่งสถานการณ์ที่จะนำมาทำการทดลองเป็น 3 สถานการณ์ด้วยกัน คือ

1. การค้นหาความซ้ำซ้อนของข้อมูลที่จะทำการจับเวลาในทุกขั้นตอนรวมถึงการสร้างตาราง และการนำเข้าข้อมูลในตารางข้อมูล ซึ่งเป็นขั้นตอนในการติดตั้งฐานข้อมูล
2. การค้นหาความซ้ำซ้อนของข้อมูลที่จะทำการจับเวลาในเฉพาะของขั้นตอนที่จะใช้ในการค้นหาความซ้ำซ้อนเท่านั้น
3. การค้นหาความซ้ำซ้อนของข้อมูลที่จะทำการจับเวลา เมื่อมีการเพิ่มข้อมูลเข้ามาใหม่ทีละข้อมูล

จะขออธิบายถึงความหมายของสถานการณ์ต่างๆ ที่จะนำมาใช้ในการทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพ สำหรับในสถานการณ์ที่ 1 ทำการลองเพื่อที่จะทำให้สามารถรู้ได้ว่าเมื่อมีการวางระบบจัดการฐานข้อมูลขึ้นมาใหม่ และเมื่อต้องทำการค้นหาความซ้ำซ้อนของข้อมูลภายในฐานข้อมูลจะต้องใช้เวลาเท่าไหร่ในการค้นหาความซ้ำซ้อนของข้อมูล สำหรับในสถานการณ์ที่ 2 จะทำการทดลองขึ้นเพื่อที่ได้รู้เวลาที่แท้จริงที่จะใช้ในการค้นหาความซ้ำซ้อนของข้อมูลว่าเมื่อไม่มีขั้นตอนในการสร้างตาราง และขั้นตอนในการนำข้อมูลเข้ามาภายในฐานข้อมูล และสำหรับในสถานการณ์ที่ 3 การค้นหาความซ้ำซ้อนของข้อมูลเมื่อมีการเพิ่มข้อมูลเข้ามาใหม่ทีละข้อมูล ทำขึ้นมาเพื่อที่จะทำให้สามารถรู้ได้ว่าเมื่อชุดข้อมูลมีการเปลี่ยนแปลงจะต้องใช้เวลาเท่าไหร่ในการค้นหาความซ้ำซ้อนของข้อมูล

จากสถานการณ์ทั้ง 3 จะต้องทำการทดลองตามแนวทางทั้ง 4 ที่ได้มีการนำเสนอไว้ก่อนหน้านี้ คือ

1 แนวทางการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมาเอง

2 แนวทางการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมาเอง และมีการใช้เทคนิคการสร้างดัชนีข้อมูลเข้าร่วมด้วย

3 แนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด

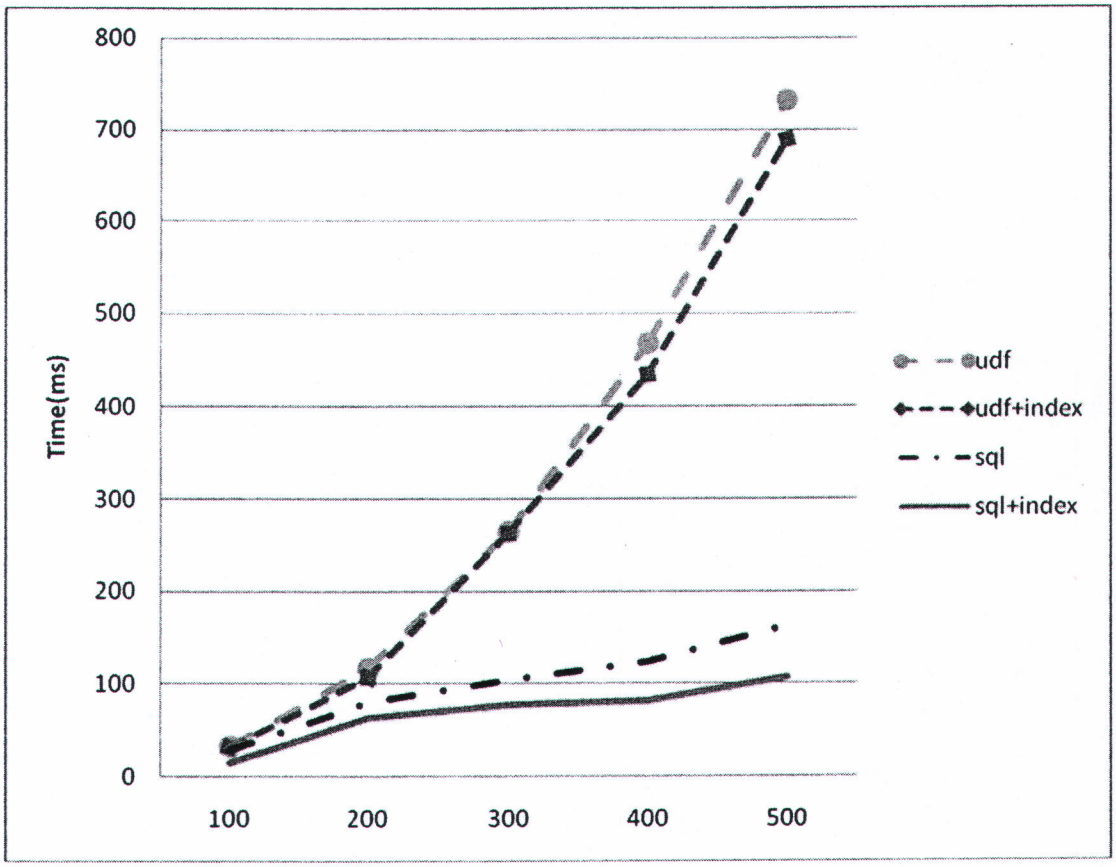
4 แนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด มีการใช้เทคนิคการสร้างดัชนีข้อมูล และเทคนิคการเขียนการสอบถามใหม่เข้าร่วมด้วย

ดังนั้นจากสถานการณ์ในหนึ่งสถานการณ์ จะต้องทำการทดลองตามแนวทางทั้ง 4 นี้ทุกครั้ง และในแต่ละแนวทางจะทำการทดลองกับขนาดที่แตกต่างกันทั้ง 5 ขนาด ซึ่งในแต่ละขนาดจะทำการทดลองทั้งหมด 10 ครั้ง ในแต่ละครั้งจะทำการเปรียบเทียบประสิทธิภาพโดยการจับเวลา และจะนำเวลาที่ได้จากการทดลองทั้ง 10 ครั้งสำหรับแต่ละขนาดมาหาค่าเฉลี่ย และจะทำอย่างนี้เรื่อยไปจนครบทุกขนาดและครบทุกแนวทาง ซึ่งจะต้องทำการทดลองทั้งหมด 5×10×4 ครั้งจึงจะได้กราฟที่จะแสดงการเปรียบเทียบในแต่ละสถานการณ์ สาเหตุที่ต้องทำการทดลองหลายครั้งแล้วจึงนำมาหาค่าเฉลี่ยก็เพราะว่าต้องการให้มีความแม่นยำมากขึ้น และในส่วนของกราฟต่างๆจะขออธิบายในหัวข้อต่อไป

4.2 ผลการทดลอง

จากการทดลองตามสถานการณ์ทั้ง 3 สถานการณ์ที่ได้นำเสนอไป ทำให้ได้ผลลัพธ์ที่จะนำมาอธิบายในรูปของกราฟแสดงประสิทธิภาพ โดยจะทำการเปรียบเทียบจากขนาดของฐานข้อมูลที่แตกต่างกันกับเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูล

4.2.1 แนวคิดในการค้นหาความซ้ำซ้อนของข้อมูล โดยจะทำการจับเวลาในทุกขั้นตอน

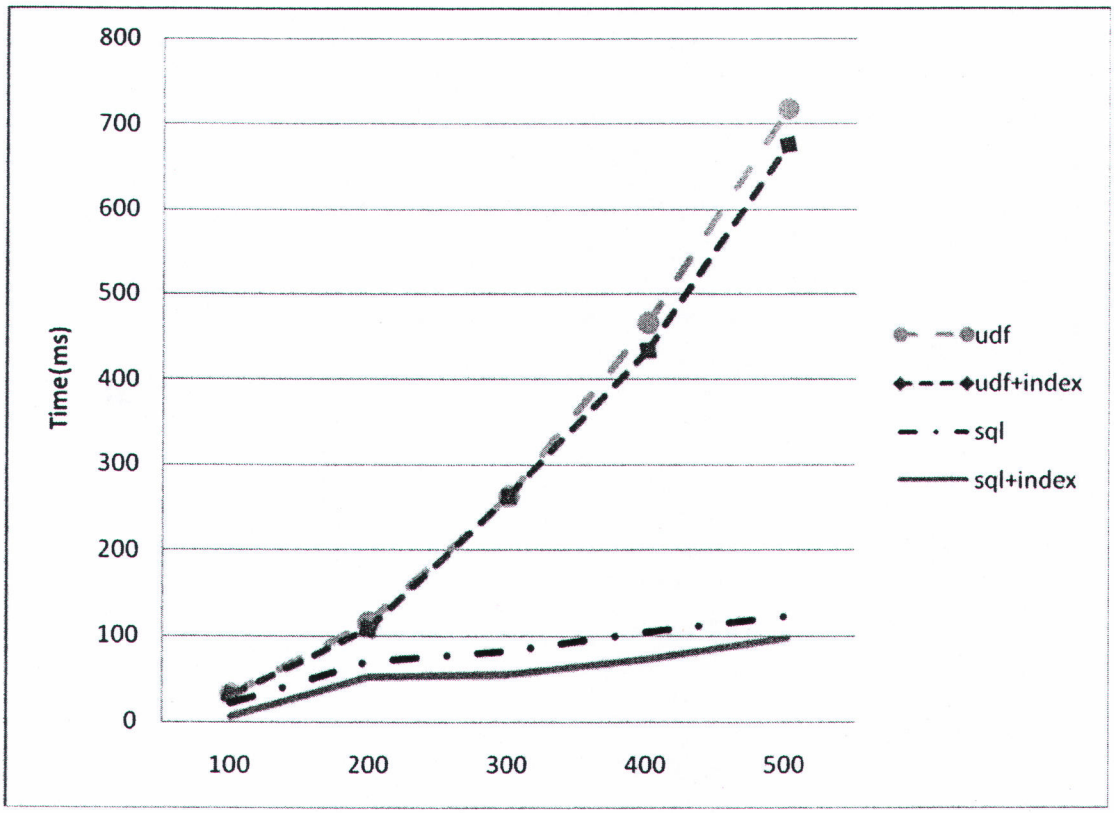


รูปที่ 4.1 แสดงกราฟเปรียบเทียบเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูลกับขนาดที่แตกต่างกันโดยจะทำการจับเวลาในทุกๆขั้นตอน

จากรูปที่ 4.1 จะแสดงเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูลโดยเปรียบเทียบกับขนาดที่แตกต่างกันถึง 5 ขนาด โดยแกนในแนวตั้งจะแสดงเวลาที่ใช้ในการค้นหาข้อมูลที่มีความซ้ำซ้อนกัน ในที่นี้ยังรวมถึงเวลาที่ใช้ในการสร้างตาราง และเวลาที่ในการนำเข้าข้อมูลทั้งหมดอีกด้วยซึ่งใช้หน่วยวัดในการเปรียบเทียบเป็นมิลลิวินาที (ms) ส่วนแกนในแนวนอนจะบอกถึงขนาดที่แตกต่างกันของฐานข้อมูลที่จะนำมาใช้ในการทดลองซึ่งมีทั้งหมด 5 ขนาด แบ่งเป็น 100, 200, 300, 400,

500 ระเบียน (Record) ตัวอย่างขั้นตอนในการค้นหาความซ้ำซ้อน เช่น ในข้อมูลขนาด 100 จะต้องใช้ฐานข้อมูลที่มีขนาด 100×100 ระเบียนเพื่อใช้ในกระบวนการค้นหาความซ้ำซ้อนกัน จากรูปที่ 4.1 จะเห็นได้ว่ามีการใช้แนวคิดทั้ง 4 แนวทางในการเปรียบเทียบประสิทธิภาพ โดยที่ให้ "udf" แทนด้วยแนวทางในการสร้างฟังก์ชันขึ้นมาเองเพื่อใช้ในการค้นหาความซ้ำซ้อน "udf+index" แทนด้วยแนวทางในการสร้างฟังก์ชันขึ้นมาเอง และมีการใช้เทคนิคในการสร้างดัชนีเพิ่มเข้ามา เพื่อใช้ในการค้นหาความซ้ำซ้อน ส่วน "sql" จะแทนด้วยแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด และในส่วนของ "sql+index" จะแทนด้วยแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด และมีการใช้เทคนิคในการสร้างดัชนีข้อมูลเพิ่มเข้ามา จากการทดลองทั้ง 5 ขนาดจะได้ผลการทดลองคือ ในขนาด 100 ระเบียน เมื่อทำการค้นหาความซ้ำซ้อนของข้อมูลจะต้องทำการเปรียบทั้งหมด 10,000 คู่ ได้ผลว่ามีความซ้ำซ้อนของข้อมูลทั้งหมด 3,960 คู่ ขนาด 200 ระเบียน จะต้องทำการเปรียบทั้งหมด 40,000 คู่ ได้ผลว่ามีความซ้ำซ้อนของข้อมูลทั้งหมด 14,276 คู่ ขนาด 300 ระเบียน จะต้องทำการเปรียบทั้งหมด 90,000 คู่ ได้ผลว่ามีความซ้ำซ้อนของข้อมูลทั้งหมด 34,830 คู่ ขนาด 400 ระเบียน จะต้องทำการเปรียบทั้งหมด 160,000 คู่ ได้ผลว่ามีความซ้ำซ้อนของข้อมูลทั้งหมด 59,492 คู่ และที่ขนาด 500 ระเบียน จะต้องทำการเปรียบทั้งหมด 250,000 คู่ ได้ผลว่ามีความซ้ำซ้อนของข้อมูลทั้งหมด 89,414 คู่ จะเห็นได้ว่าแนวทางในการสร้างฟังก์ชันขึ้นมาเอง และแนวทางในการสร้างฟังก์ชันขึ้นมาเองร่วมกับการเพิ่มเทคนิคในการสร้างดัชนีข้อมูลเข้ามา ให้ผลลัพธ์ที่ไม่แตกต่างกันอย่างชัดเจน แต่ในแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด และแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมดร่วมกับการเพิ่มเทคนิคในการสร้างดัชนีข้อมูล จะให้ผลลัพธ์ในการค้นหาความซ้ำซ้อนของข้อมูลที่มีความรวดเร็วกว่าแนวทางทั้ง 2 แนวทางก่อนหน้านี้อย่างชัดเจน และแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมดร่วมกับการเพิ่มเทคนิคในการสร้างดัชนีข้อมูลเข้ามาจะเป็นแนวทางที่ให้ผลลัพธ์ที่มีความเร็วที่สุด ในการค้นหาความซ้ำซ้อนของข้อมูล และมีแนวโน้มว่าจะมีความเร็วเพิ่มมากขึ้นกว่าแนวทางอื่นเมื่อฐานข้อมูลมีขนาดใหญ่ขึ้น

4.2.2 แนวคิดในการค้นหาความซ้ำซ้อนของข้อมูล โดยจะทำการจับเวลาเฉพาะขั้นตอนการค้นหาความซ้ำซ้อนเท่านั้น

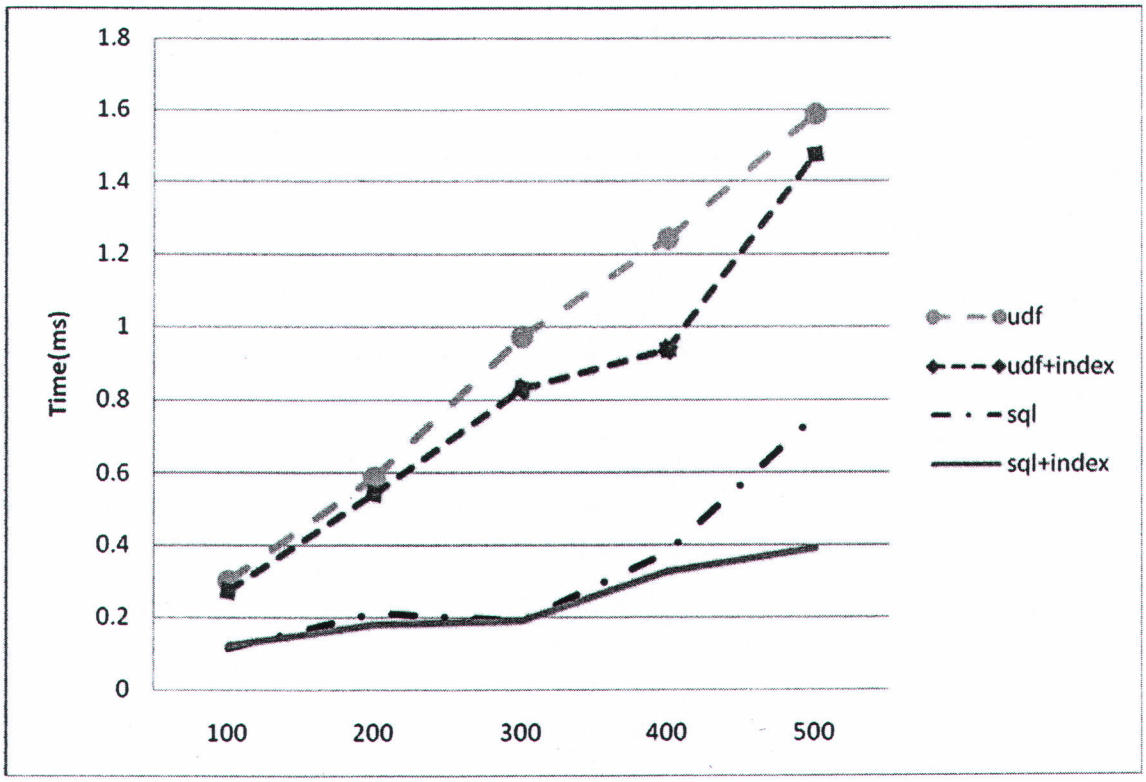


รูปที่ 4.2 แสดงกราฟเปรียบเทียบเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูลกับขนาดที่แตกต่างกันโดยจะทำการจับเวลาเฉพาะขั้นตอนในการค้นหาความซ้ำซ้อนเท่านั้น

จากรูปที่ 4.2 จะแสดงเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูลกับขนาดที่แตกต่างกันโดยจะทำการจับเวลาเฉพาะขั้นตอนในการค้นหาความซ้ำซ้อนเท่านั้นซึ่งจะใช้ฐานข้อมูลเดียวกันกับการทดลองในสถานการณ์แรก ดังนั้นจำนวนของการเปรียบเทียบ รวมถึงจำนวนของคู่ที่จะได้จากกระบวนการในการค้นหาความซ้ำซ้อนจะมีจำนวนเท่ากับในสถานการณ์แรกทุกประการ เพียงแต่ในการทดลองลองนี้จะไม่ทำการจับเวลาในส่วนของขั้นตอนการสร้างตาราง และการนำเข้าข้อมูลในฐานข้อมูล เพื่อที่จะทำให้ได้ทราบเวลาที่ใช้งานจริงๆ สำหรับการค้นหาความซ้ำซ้อนของข้อมูล เนื่องจากในความเป็นจริงแล้ว กระบวนการในการค้นหาความซ้ำซ้อนของข้อมูลไม่จำเป็นต้องมีการสร้างตารางใหม่ทุกครั้ง ดังนั้นจึงได้ทำการทดลองสำหรับสถานการณ์นี้ขึ้นมาเพื่อให้สามารถที่

จะประมาณเวลาที่จะใช้การค้นหาความซ้ำซ้อนของข้อมูล เมื่อนำไปใช้งานจริงๆได้ จากกราฟจะเห็นได้ว่า แนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมดร่วมกับการเพิ่มเทคนิคในการสร้างดัชนีข้อมูลยังคงเป็นแนวทางที่ให้ผลลัพธ์ที่มีความรวดเร็วที่สุดเหมือนกับในสถานการณ์ก่อนหน้านี้

4.2.3 แนวคิดในการค้นหาความซ้ำซ้อนของข้อมูล โดยจะทำการจับเวลาเมื่อฐานข้อมูลมีการเปลี่ยนแปลง



รูปที่ 4.3 แสดงกราฟเปรียบเทียบเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูลกับขนาดที่แตกต่างกันโดยจะทำการจับเวลาเมื่อมีการเปลี่ยนแปลงของข้อมูลภายในฐานข้อมูล

จากรูปที่ 4.3 จะแสดงเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูลกับขนาดที่แตกต่างกันโดยจะทำการจับเวลาเมื่อฐานข้อมูลมีการเปลี่ยนแปลงทีละ 1 ระเบียน ซึ่งกระบวนการในการค้นหาความซ้ำซ้อนจะมีความแตกต่างกับในสถานการณ์ที่ 1 และสถานการณ์ที่ 2 อย่างชัดเจน โดยในสถานการณ์นี้จะทำการค้นหาความซ้ำซ้อนของข้อมูลในขนาดที่แตกต่างกัน 5 ขนาด คือ 100, 200, 300, 400, 500 ในส่วนของข้อมูลขนาด 100 จะทำการเปรียบเทียบทั้งหมด 100×1 ระเบียน ซึ่ง

สามารถคิดเป็สมการการได้ดังนี้ ขนาดของฐานข้อมูล×1 สาเหตุที่มีการทดลองในสถานการณ์นี้เพราะว่าในความเป็นจริงสถานการณ์ในการเพิ่มของฐานข้อมูลทีละ 1 ระเบียนมีความเป็นไปได้ที่จะเกิดขึ้นบ่อยๆ ในระบบจัดการสาระสนเทศต่างๆ เช่น ระบบการสมัครสมาชิกเพื่อใช้ระบบสาระสนเทศต่างๆ ดังนั้นจึงมีการทดลองในสถานการณ์นี้เพื่อให้สามารถที่จะประมาณเวลาที่จะใช้ในการค้นหาความซ้ำซ้อนในโลกของความเป็นได้ และทำให้สามารถที่จะทราบได้ว่ากระบวนการในการค้นหาความซ้ำซ้อนของข้อมูลภายในฐานข้อมูลที่ผู้วิจัยได้นำเสนอ จะสามารถนำไปใช้งานจริงได้หรือไม่ ซึ่งจากผลการทดลองจะเห็นได้อย่างชัดเจนว่าแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมด จะให้ผลลัพธ์ในการค้นหาความซ้ำซ้อนของข้อมูลที่เร็วกว่าแนวทางที่ใช้การสร้างฟังก์ชันขึ้นมาเอง และแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมดร่วมกับการใช้ดัชนีข้อมูล เป็นแนวทางที่ให้ผลลัพธ์ที่มีความรวดเร็วที่สุดเหมือนกับในสถานการณ์ก่อนหน้านี้ อีกทั้งยังมีแนวโน้มว่าจะมีความเร็วกว่าแนวทางอื่นๆอย่างเห็นได้ชัดเมื่อฐานข้อมูลมีขนาดใหญ่ขึ้น

4.3 สรุปผลการทดลอง

จากการทดลองทั้ง 3 สถานการณ์จะเห็นได้ว่าแนวทางการอิมพลีเมนต์ด้วยภาษาพีแอลเอสคิวแอลทั้งหมดร่วมกับเทคนิคการใช้ดัชนีข้อมูลเข้ามา จะให้ผลลัพธ์ที่รวดเร็วที่สุดทั้ง 3 สถานการณ์ และยังมีแนวโน้มว่าจะมีความรวดเร็วมากขึ้นในการค้นหาความซ้ำซ้อนของข้อมูลกว่าในแนวทางอื่นๆ เมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่ขึ้น