บทที่ 1

บทนำ

1.1 ที่มาและปัญหาของการศึกษา

ความซ้ำซ้อนของข้อมูล คือ การที่ระบบสารสนเทศทำการเก็บข้อมูลมากกว่าหนึ่งชุดข้อมูลที่เหมือนกัน ซึ่งอาจเป็นข้อมูลจากเอนทิตี (Entity) เดียวกันในโลกแห่งความเป็นจริง แต่ระบบสารสนเทศไม่สามารถที่จะรับรู้ได้ว่าเป็นข้อมูลเหล่านั้นเป็นข้อมูลชุดเดียวกัน

ในระบบสารสนเทศขนาดใหญ่ อาจมีความซ้ำซ้อนของข้อมูลเกิดขึ้นได้ในหลายลักษณะ เช่น ในการสมัครสมาชิกของผู้ใช้บริการหนึ่งๆอาจมีการสมัครซ้ำได้ เนื่องจากผู้ใช้บริการ อาจลืมว่าได้ทำการสมัครสมาชิกไปแล้ว หรือในกรณีที่มีการสมัครสมาชิกมากกว่าหนึ่งสาขา หรือการที่บุคคลมีชื่อ หรือนามสกุลที่เหมือนกันทำให้ไม่สามารถที่จะยืนยันได้ว่าเป็นบุคคลเดียวกันหรือไม่จะทำให้เกิดความซ้ำซ้อนของข้อมูลในฐานข้อมูล และทำให้การยืนยันว่าเป็นบุคคลคนเดียวกันมีความผิดพลาด

การยืนยันข้อมูลที่ผิดพลาดอาจจะทำให้เกิดความเสียหายให้กับตัวบุคคลได้ในหลายๆ รูปแบบ และมีผลกระทบที่แตกต่างกันออกไปในแต่ละกรณี เช่น การยืนยันข้อมูลที่ผิดพลาดของข้อมูลที่เกี่ยวข้องกับธุรกรรมทางการเงิน อาจทำให้เกิดความเสียหายในด้านทรัพย์สินได้ หรือกรณีของความผิดพลาดในการยืนยันข้อมูลที่จำมาใช้ในการอ้างสิทธิต่างๆ เช่น สิทธิในการเลือกตั้ง หรือในการสมัครสมาชิกของระบบสารสนเทศที่ต้องการความเป็นส่วนตัว เช่น เบอร์โทรศัพท์ จดหมายอิเล็กทรอนิกส์ (Email) และยังมีอีกหลายกรณีที่ไม่ได้ยกตัวอย่างขึ้นมา โดยในโครงร่างวิทยานิพนธ์ฉบับนี้จะสนใจถึงปัญหาที่อาจจะมีผลกระทบกับตัวบุคคลเท่านั้น

การหาความซ้ำซ้อนของข้อมูลในระบบสารสนเทศต่างๆ เป็นกระบวนการที่มีความซับซ้อนมาก เนื่องจากต้องทำการเปรียบเทียบข้อมูลในทุกๆความสัมพันธ์ (Relation) ซึ่งอาจเป็นไป

ได้ว่า ข้อมูลที่ซ้ำซ้อนกันที่มีสาเหตุมาจากกรณีที่อธิบายไว้ข้างบน จะขออธิบายถึงรายละเอียดต่างๆ ที่อาจเกิดได้เป็นกรณีไป 1) การสะกดคำผิดในขั้นตอนการบันทึกข้อมูลครั้งแรก เช่น ลูกค้าที่เข้ามาสมัครสมาชิกในระบบสารสนเทศ ชื่อ "เกียรติศักดิ์ จันทร์หอม" แต่ได้ทำการสมัครสมาชิกให้ในชื่อ "เกียรติศักดิ์ จันหอม" ไปเมื่อต้องการทำการค้นหาข้อมูลของลูกค้าที่ชื่อ "เกียรติศักดิ์ จันทร์หอม" จะทำให้ค้นหาข้อมูลของลูกค้าในระบบสารสนเทศไม่พบ ดังนั้นลูกค้าอาจทำการสมัครสมาชิกใหม่ในชื่อของ "เกียรติศักดิ์ จันทร์หอม" ซึ่งที่จริงแล้วระบบมีข้อมูลของลูกค้าคนนี้อยู่แล้ว เพียงแต่ถูกเก็บข้อมูลไว้ในชื่อในอีกชื่อหนึ่ง ในกรณีนี้จะทำให้เกิดความซ้ำซ้อนของข้อมูลขึ้น 2) กรณีของการเปลี่ยนแปลงข้อมูล เช่น ลูกค้าที่ชื่อ "สมใจ ขยันยิ่ง" แต่ภายหลังได้ทำการเปลี่ยนนามสกุลจาก "ขยันยิ่ง" เป็น "นอนทั้งวัน" เพราะได้แต่งงานกับ "สมคิด นอนทั้งวัน" ทำให้ข้อมูลในระบบสารสนเทศที่เคยใช้ในชื่อของ "สมใจ ขยันยิ่ง" ไม่สามารถใช้ได้เนื่องจากมีนามสกุลไม่ตรงกัน ทำให้ไม่สามารถยืนยันตัวบุคคลได้ว่าเป็นคนเดียวกัน จากกรณีนี้ลูกค้าอาจจะสมัครสมาชิกของระบบสารสนเทศนั้นใหม่ในชื่อของ "สมใจ นอนทั้งวัน" และได้มีการเปลี่ยนที่อยู่ใหม่แต่ข้อมูลในด้านอื่นๆ เช่น ข้อมูลของวันเกิด บัตรประชาชน และเบอร์โทรศัพท์ยังเป็นข้อมูลที่เหมือนกับข้อมูลที่อยู่ในชื่อของ "สมใจ ขยันยิ่ง" จะทำให้เกิดความซ้ำซ้อนของข้อมูลทั้งที่เป็นบุคคลเดียวกันแต่ทำการเปลี่ยนนามสกุล และที่อยู่ใหม่ เนื่องจากระบบสารสนเทศนั้นไม่สามารถที่จะยืนยันได้ว่าเป็นบุคคลเดียวกัน 3) กรณีการขาดหายไปของข้อมูลในการบันทึกครั้งแรก เช่น การบันทึกข้อมูลที่ไม่ครบในทุกๆแอตทริบิวต์ (Attribute) โดยปล่อยไว้เป็นค่าว่างๆ เมื่อระบบทำการค้นหาความซ้ำซ้อนของข้อมูลจะทำให้พบความซ้ำซ้อนของข้อมูลที่ไม่ถูกต้องได้ เช่น ถ้าไม่ได้ทำการบันทึกข้อมูลด้านที่อยู่ปกติระบบจะทำการบันทึกให้ว่าเป็นค่าว่างๆ หรือไม่รู้ (Unknown) เมื่อมีบุคคลอื่นที่ไม่ได้ทำการบันทึกข้อมูลด้านที่อยู่ลงไปเหมือนกัน จะทำให้การตรวจสอบความซ้ำซ้อนของข้อมูลในระบบสารสนเทศเกิดความผิดพลาด เพราะความเหมือนกันของข้อมูลในด้านที่อยู่

จากปัญหาการตรวจสอบความซ้ำซ้อนกันของข้อมูลที่ได้กล่าวไปแล้ว ได้มีการเสนอแนวทางการแก้ปัญหาไว้ใน [1] โดยมีขั้นตอนดังนี้

1) การเลือกเดสคริปชัน (Description) เพื่อใช้ในการกำหนดแอตทริบิวต์ที่จะนำไปใช้ในการจำแนก

2) การจำแนก (Classification) ของข้อมูลที่มีความเหมือนกัน

จากแนวทางในการแก้ปัญหาการตรวจสอบการซ้ำซ้อนกันของข้อมูลที่กล่าวมาแล้ว ทางเลือกหนึ่งในการอิมพลีเมนต์ คือ การอิมพลีเมนต์ด้วยภาษาระดับสูงเพื่อสร้างโมดุลภายนอกสำหรับการตรวจสอบ ซึ่งโมดุลนี้จะทำงานร่วมกับระบบจัดการฐานข้อมูลด้วยการนำข้อมูลเข้า (Import) และการส่งข้อมูลออก (Export) ซึ่งการอิมพลีเมนต์ในลักษณะนี้ อาจไม่ใช่แนวทางที่มีประสิทธิภาพ หากข้อมูลในฐานข้อมูลมีความเปลี่ยนแปลงบ่อยครั้ง อีกทั้งแนวทางดังกล่าวยังไม่ได้ใช้กลไกในการปรับปรุงประสิทธิภาพการสอบถาม (Query Optimization) ที่ถือได้ว่าเป็นคุณสมบัติเด่นประการหนึ่งของระบบจัดการฐานข้อมูล

1.2 แนวทางการแก้ปัญหา

ใช้การอิมพลีเมนต์บนระบบจัดการฐานข้อมูลโดยภาษาเอสคิวแอลทั้งหมด เนื่องจากต้องการที่จะแก้ไขปัญหาการนำเข้า และการส่งออกของข้อมูล เพราะว่าการใช้ภาษาเอสคิวแอลในการอิมพลีเมนต์ ทำให้สามารถที่จะทำการค้นหาความซ้ำซ้อนข้อมูลได้เลยภายในระบบจัดการฐานข้อมูล เพื่อลดขั้นตอนที่ยุ่งยากในกระบวนนำเข้าข้อมูล และการส่งออกข้อมูลที่เคยใช้ในการค้นหาความซ้ำซ้อนในโมดุลแบบเก่า เพื่อให้สามารถที่จะนำไปใช้งานได้จริง

โดยในงานวิจัยนี้จะมีการพัฒนาแนวทางในการค้นหาความซ้ำซ้อนของข้อมูล ซึ่งทำงานบนระบบจัดการฐานข้อมูล ให้สามารถประมวลผลได้อย่างมีประสิทธิภาพ โดยใช้เทคนิคที่สามารถทำได้บนระบบจัดการฐานข้อมูล ซึ่งได้แก่ การเขียนการสอบถามใหม่ การทำดัชนี และการดีนอร์มอลไลเซชัน

1.3 วัตถุประสงค์ของการศึกษา

พัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลซึ่งทำงานบนระบบจัดการฐานข้อมูล ให้สามารถประมวลผลได้อย่างมีประสิทธิภาพ

1.4 ประโยชน์ที่ได้รับจากการศึกษา เชิงทฤษฎี และ/หรือ เชิงประยุกต์

สามารถนำแนวทางที่พัฒนาขึ้นไปประยุกต์ใช้ค้นหาความซ้ำซ้อนในองค์กรทั่วไปได้

1.5 ขอบเขตการทำวิจัย

1.5.1 ใช้ระบบจัดการฐานข้อมูลออราเคิล (Oracle) และฐานข้อมูล T-CPH Benchmark

1.5.2 วัดประสิทธิภาพจากเวลาที่ใช้ในการทำงานจริง โดยเปรียบเทียบระหว่างแนวทางที่พัฒนาขึ้น ซึ่งประยุกต์ใช้เทคนิคการเขียนการสอบถามใหม่ การสร้างดัชนี และการนอร์มอลไลเซชัน/ดีนอร์มอลไลเซชัน กับแนวทางที่ไม่ได้ใช้เทคนิคดังกล่าว

1.6 วิธีการทำวิจัย

1.6.1 ศึกษาเอกสารข้อมูลและงานวิจัยที่เกี่ยวข้อง

1.6.2 ศึกษาทฤษฎีที่เกี่ยวข้อง

1.6.3 ออกแบบและพัฒนาแบบจำลอง

1.6.4 ออกแบบและสร้างขั้นตอนวิธีการสร้างดัชนี

1.6.5 ออกแบบและสร้างขั้นตอนวิธีการดีนอร์มอลไลเซชชัน

1.6.6 ออกแบบและปรับปรุงประสิทธิ์ภาพแบบสอบถาม

1.6.7 สร้างโมดุลตามที่ได้ออกแบบไว้

1.6.8 วิเคราะห์และสรุปผล จัดทำและเสนอรายงานวิทยานิพนธ์

1.7 เครื่องมือที่ใช้ในการพัฒนา

ในงานวิจัยนี้มีเครื่องมือในการพัฒนาดังต่อไปนี้

1.7.1 ฮาร์ดแวร์

- 2.66 GHz Intel Core i5 4GB main memory

1.7.2 ซอฟต์แวร์

- Microsoft Window 7
- Oracle database 11g R2