ความสัมพันธ์ระบบคลังข้อมูลกับระบบฐานข้อมูล

ในปัจจุบันมีการใช้ฐานข้อมูลอย่างกว้างขวางในระบบงานทั่วไป จึงมีการวิจัยและพัฒนาวิธีเก็บข้อมูลจำนวนมาก รวมถึงการค้นหาและนำข้อมูลที่ต้องการออกมาจากระบบฐานข้อมูลด้วย แต่เนื่องจากระบบฐานข้อมูลทั่วไป (Operational Database) ที่นิยมใช้อยู่ในปัจจุบันมีหลักในการเก็บข้อมูลที่เน้นในเรื่องการลดความซ้ำซ้อน (redundancy) รักษาความถูกต้อง (integrity) ลดการสูญหายของข้อมูล (information lost) และลดความผิดพลาดที่เกิดขึ้นจากการแก้ไขข้อมูล (Update Anomalies)

เนื่องจากฐานข้อมูลทั่วไป (Operational Database) มีลักษณะดังได้กล่าวมาแล้วจึงมีความสามารถเพียงแค่การเรียกใช้ข้อมูลที่มีอยู่ แต่ไม่สามารถจะนำมาช่วยในการสนับสนุนการตัดสินใจได้ เพราะเมื่อมีการเรียกใช้ข้อมูลจะต้องเรียกใช้ข้อมูลจากฐานข้อมูลขนาดใหญ่ ซึ่งมีข้อมูลจำนวนมหาศาลและมีการแตกตารางที่นอร์มัลไลซ์ (normalized table) แล้วออกเป็นหลายตาราง จึงไม่รองรับคำถามที่ต้องการจะนำมาใช้ช่วยในการสนับสนุนการตัดสินใจ (decision support queries) มีการรวม (join) กันของตารางต่างๆที่ซับซ้อน ซึ่งจะทำให้มีประสิทธิภาพของการค้นหาข้อมูลจากฐานข้อมูลน้อยลง และทำงานช้าลง ไม่สามารถเรียกใช้ข้อมูลที่ต้องการได้ทั้งหมดเพราะมีรูทีนอัตโนมัติ (Automate Routine) จึงมีความสามารถในการค้นหาข้อมูลแบบที่ไม่ซ้ำซ้อนเท่านั้น นอกจากนี้การเก็บข้อมูลในระบบฐานข้อมูลทั่วไป (Operational Database) ยังไม่มีการเก็บข้อมูลย้อนหลัง (historical data) เพื่อใช้ช่วยในการคาดคะเนแนวโน้มที่คาดว่าจะเป็นไปได้ในอนาคต

ดังนั้นระบบคลังข้อมูลจึงได้ถูกคิดขึ้นมาเพื่อช่วยให้ผู้ใช้เรียกใช้ข้อมูลที่มีอยู่ได้อย่างมีประสิทธิภาพสูงสุดด้วยวิธีที่สร้างสรรค์เพราะธรรมชาติที่แตกต่างกันระหว่างระบบฐานข้อมูลคลังข้อมูลและระบบฐานข้อมูลทั่วไป ดังนั้นฐานข้อมูลคลังข้อมูลจะต้องมีคุณสมบัติดังนี้

1. Subject Oriented ข้อมูลจะต้องถูกสร้างขึ้นจากหัวข้อ (subject) ธุรกิจที่สนใจ เช่น ถ้าบริษัทประกันภัยต้องการใช้คลังข้อมูล ฐานข้อมูลที่ได้จะต้องสร้างขึ้นจากประวัติลูกค้า, เบี้ยประกัน และการเรียกร้องแทนที่จะแยกตามชนิดของผลิตภัณฑ์ หรือบริการประกันภัย/ประกันชีวิต ข้อมูลที่สร้างขึ้นจะประกอบด้วยหัวข้อที่เก็บเฉพาะข่าวสารที่จำเป็น สำหรับกระบวนการตัดสินเท่านั้น
2. Integrated ข้อมูลถูกรวบรวมจากแหล่งต่างๆ จากระบบปฏิบัติการ, รูปแบบของข้อมูล, แพลตฟอร์มที่หลากหลาย สร้างขึ้นเป็นฐานข้อมูลที่สอดคล้องเป็นหนึ่งเดียว เช่นค่าของตัวแปรตัวเดียวในแต่ละฐานข้อมูลอาจต่างกัน ฐานข้อมูลหนึ่งอาจใช้ 0 และ 1 อีกฐานข้อมูลหนึ่งอาจใช้ T และ F ดังนั้นฐานข้อมูลที่สร้างใหม่จะต้องได้รับการกำหนดค่าตัวแปรให้เหมือนกันเป็นหนึ่งเดียว
3. Time-variant ข้อมูลซึ่งใช้ตัดสินใจที่เก็บไว้จะต้องมีอายุประมาณ 5 ถึง 10 ปี เพื่อใช้เปรียบเทียบ หาแนวโน้ม และทำนายผลลัพธ์ในอนาคตได้
4. Non-volatile ข้อมูลจะไม่อัพเดตหรือถูกทำให้เปลี่ยนแปลงง่ายๆ ผู้ใช้สามารถใช้ฐานข้อมูลคลังข้อมูลได้เพียงแค่โหลดและเข้าถึงเท่านั้น

โดยระบบฐานข้อมูลคลังข้อมูลจะแยกกลุ่มข้อมูลสารสนเทศที่ใช้ในการวิเคราะห์ทางธุรกิจออกจากฐานข้อมูลที่ใช้ประจำวัน (Operational Database) มาเก็บอยู่ในระบบจัดการฐานข้อมูล (Relational Database Management Systems) ประสิทธิภาพสูงสุด และทำให้การเรียกใช้ข้อมูลชุดนี้ทำได้อย่างยืดหยุ่น จากเครื่องมือที่อยู่บนเครื่องคอมพิวเตอร์เดสก์ทอปทั่วไป โดยลด off-loading เพิ่มกลไกการช่วยตัดสินใจ ปรับปรุงเวลาที่ตอบสนอง (response time) รวดเร็วขึ้นอย่างมากและผู้บริหารสามารถเรียกข้อมูลรายละเอียดที่จำเป็นที่ถูกเก็บมาก่อนหน้านี้ (historical data) มาใช้ช่วยในการตัดสินใจทางธุรกิจแม่นยำขึ้น

ความแตกต่างอีกประการหนึ่งก็คือผู้ใช้คลังข้อมูลมักจะต้องการจัดกลุ่มข้อมูลด้วยตนเองมากกว่าผู้ใช้ในระบบฐานข้อมูลธรรมดา ยกตัวอย่างผู้ใช้อาจต้องการวิเคราะห์ผลกระทบของการทำการตลาดแบบต่างๆ อาจต้องการจัดกลุ่มการขายสินค้าแยกตามผลิตภัณฑ์ หรือรูปแบบของการจัดผลิตภัณฑ์ เช่น การห่อรวมสินค้าไว้ในบรรจุภัณฑ์สีต่างๆ หรือการรวมผลิตภัณฑ์ต่างรูปแบบไว้ด้วยกัน ในกรณีต่างๆ เหล่านี้ผู้ใช้ต้องการที่จะเลือกจัดกลุ่มข้อมูลได้ตามใจชอบ นอกจากการนำข้อมูลเข้ามารวมกันแล้ว ผู้ใช้ยังอาจต้องการที่จะแยกแยะข้อมูลในแบบที่ตนเองต้องการได้ ยกตัวอย่างในการจัดทำคลังข้อมูลเกี่ยวกับนักวิจัยและผลงานวิจัยของประเทศ หน่วยงานอาจจัดเก็บข้อมูลเอาไว้เป็นกลุ่มก้อนโดยไม่ได้แยกสาขา แต่ต่อมาผู้ใช้อาจต้องการนำข้อมูลนักวิจัยมาวิเคราะห์แยกแยะว่าทั้งประเทศมีนักวิจัยสาขาต่างๆ เป็นจำนวนเท่าใด ทำงานวิจัยด้านใดบ้าง ใช้เงินด้านวิจัยไปเท่าใด เป็นต้น โดยปกติแล้วการจัดทำฐานข้อมูลให้สามารถวิเคราะห์แยกแยะข้อมูลในแบบนี้ได้นั้นเป็นเรื่องไม่ยาก แต่ในการออกแบบคลังข้อมูลนั้นจำเป็นต้องเผื่อให้ผู้ใช้หลายคนสามารถแยกแยะข้อมูลตามความต้องการที่แตกต่างกันได้ด้วย ผู้ใช้จำนวนมากในปัจจุบันนี้อาจใช้ซอฟต์แวร์หลากหลายประเภทสำหรับเครื่องคอมพิวเตอร์ส่วนบุคคล ผู้ใช้บางคนอาจจะใช้โปรแกรมสเปรดชีตในการวิเคราะห์ข้อมูล และผู้ใช้บางคนอาจต้องการใช้โปรแกรมวิเคราะห์สถิติอื่นๆ ดังนั้นผู้ใช้เหล่านี้อาจจะมีความต้องการในการนำเข้าข้อมูลจากคลังข้อมูลมาไว้ในแฟ้มข้อมูลที่มีรูปแบบตรงกับโปรแกรมที่ตนต้องการใช้ ความต้องการด้านนี้นับว่าสำคัญมากที่สุดในการจัดทำคลังข้อมูล

งานอย่างหนึ่งที่นิยมใช้ฐานข้อมูลกันมากก็คืองานบันทึกข้อมูลธุรกรรมเอาไว้เพื่อประมวลผล ข้อมูลธุรกรรมเหล่านี้ได้แก่ ข้อมูลการสั่งซื้อสินค้าของลูกค้า ข้อมูลการซื้อบัตรโดยสารเครื่องบิน ข้อมูลการฝากหรือถอนเงินของลูกค้าธนาคาร แต่เดิมนั้นการบันทึกข้อมูลธุรกรรมเริ่มต้นด้วยการใช้กระดาษแบบฟอร์มสำหรับให้ลูกค้ากรอกข้อมูล จากนั้นจึงนำแบบฟอร์มมาบันทึกข้อมูลลงในฐานข้อมูลของระบบคอมพิวเตอร์ในแบบแบตช์ (batch) ปัจจุบันนี้การบันทึกข้อมูลธุรกรรมได้เปลี่ยนไปเป็นระบบออนไลน์ (online) เป็นส่วนใหญ่ ในระบบแบบนี้กระบวนการบันทึกข้อมูลมีลักษณะอัตโนมัติมากขึ้นและใช้อุปกรณ์บันทึกข้อมูลที่สามารถเก็บข้อมูลลงในฐานข้อมูลของระบบคอมพิวเตอร์ได้ทันที เช่น การใช้อุปกรณ์ฝากถอนเงินโดยอัตโนมัติ (ATM) ทำให้สามารถประมวลผลการฝากถอนเงินและบันทึกข้อมูลที่เกิดขึ้นได้ทันที หรือในห้างสรรพสินค้าก็มีการใช้เครื่องบริการ ณ จุดขาย (Point of Sale; POS) สำหรับอ่านรหัสแท่ง แสดงราคาสินค้า แล้วบันทึกข้อมูลการขายไป

เก็บไว้ในฐานข้อมูลได้ทันที การดำเนินการในลักษณะนี้เรียกกันว่า *การประมวลผลธุรกรรมออนไลน์ (On-Line Transaction Processing; OLTP)*

1. ลักษณะงานการประมวลผลธุรกรรมออนไลน์และการประมวลผลเชิงวิเคราะห์ออนไลน์

ระบบ OLTP โดยทั่วไปจะต้องสามารถดำเนินการกับข้อมูลธุรกรรมได้อย่างมีประสิทธิภาพ งานที่ทำกับข้อมูลได้แก่การปรับค่าของข้อมูลให้เป็นปัจจุบันและการเพิ่มข้อมูลลงไปในฐานข้อมูล ข้อมูลเหล่านี้อาจจะมีจำนวนมากและเพิ่มขึ้นตลอดเวลา ณ เวลาใดเวลาหนึ่งอาจจะมีการประมวลผลข้อมูลจำนวนนับแสนเรคอร์ดได้ เช่น ณ สนามบินแต่ละแห่งจะมีผู้โดยสารเข้ามารับบัตรที่นั่งของสายการบินต่างๆ เป็นจำนวนนับหมื่นๆ คน คอมพิวเตอร์ของสายการบินจะต้องตรวจสอบการสำรองที่นั่ง ต้องบันทึกเลขที่นั่งและเที่ยวบินรวมทั้งอาจจะต้องปรับเปลี่ยนโยกย้ายข้อมูลจากเที่ยวบินหนึ่งไปอีกเที่ยวบินหนึ่งได้ด้วย หรือในกรณีของศูนย์การค้า และซูเปอร์มาร์เก็ต จะมีการบันทึกเรคอร์ดการขายเพิ่มเข้าไปในฐานข้อมูลการขายตลอดเวลา รวมแล้ววันละเป็นหมื่นๆ รายการ การออกแบบระบบ OLTP แบบนี้จำเป็นต้องหาทางให้ระบบสามารถทำงานได้อย่างถูกต้องรวดเร็วตลอดเวลา เอื้ออำนวยให้ผู้ใช้จำนวนมากสามารถใช้ระบบได้พร้อมกัน อีกทั้งยังต้องสามารถแก้ไขฟื้นสภาพให้กลับดีดังเดิมได้หากเกิดความขัดข้องเสียหาย

การที่จะจัดทำระบบ OLTP ให้มีความสามารถในแบบนี้ได้ต้องคำนึงถึงปัจจัยต่อไปนี้

1) ขนาดและตำแหน่งของ rollback segment

2) ดัชนี การจัดกลุ่ม และ การคำนวณตำแหน่งที่อยู่ (hashing)

3) การออกแบบข้อมูลธุรกรรมให้เหมาะกับงานประยุกต์

4) หน่วยเก็บและเนื้อที่ว่างสำหรับการเก็บข้อมูลใหม่

5) ความเข้าใจลักษณะงานประยุกต์และการเขียนคำสั่งสำหรับค้นคืนข้อมูล

6) การปรับปรุงสมรรถนะของระบบอย่างต่อเนื่อง

ระบบ OLTP ที่พัฒนาขึ้นโดยใช้เทคนิคด้านฐานข้อมูลตามปกติมักจะไม่สามารถรับกับปริมาณข้อมูลที่เพิ่มขึ้นอย่างมากมายเป็นประจำทุกวันได้ การนำระบบเช่นนี้มาใช้จึงมีความเสี่ยงที่จะเกิดความผิดพลาดเสียหายขึ้น วิธีการแก้ไขก็คือการแยกฐานข้อมูลออกมาเป็นส่วน ๆ ให้เหมาะกับการใช้งาน

งานที่เกี่ยวข้องกับฐานข้อมูลอีกอย่างหนึ่งก็คืองานที่เรียกว่า *การประมวลผลเชิงวิเคราะห์ออนไลน์ (On-Line Analytical; OLAP)* ระบบ OLTP ที่กล่าวไปแล้วนั้นเน้นที่การบันทึกเก็บข้อมูลใหม่ๆ เพิ่มเข้าไว้ในฐานข้อมูล

ส่วนระบบ OLAP นั้นเน้นที่การค้นคืนข้อมูลที่มีอยู่แล้วจากฐานข้อมูลเพื่อนำมาวิเคราะห์อย่างละเอียด ผู้ใช้ระบบ OLAP ส่วนใหญ่คือผู้บริหาร นักวิจัยตลาด นักสถิติ หรือ ผู้ใช้อื่นๆ ดังนั้นปัจจัยสำคัญสำหรับความสำเร็จของระบบ OLAP ก็คือระบบจะต้องทำงานได้รวดเร็ว สามารถค้นหาข้อมูลจากฐานข้อมูลขนาดใหญ่มาคำนวณได้อย่างครบถ้วนไม่ตกหล่น ในขณะเดียวกันระบบก็จะต้องมีความมั่นคง ไม่ผิดพลาดได้ง่ายระหว่างการใช้งาน

ปัจจัยที่จะทำให้ได้ตามที่กล่าวนี้มีอยู่สามข้อคือ

1) จะต้องมีระบบจัดคำสั่งค้นคืนข้อมูลให้ทำงานได้รวดเร็วที่สุด (query optimization)

2) การจัดดัชนี จัดกลุ่มข้อมูล และ การคำนวณตำแหน่งที่อยู่ข้อมูล

3) การประมวลผลคำสั่งค้นคืนในแบบขนาน โดยเฉพาะเมื่อใช้หน่วยเก็บแบบ RAID

แม้ว่าระบบ OLTP และ OLAP นี้จะเกี่ยวข้องกับข้อมูลธุรกรรมเหมือนกันแต่ก็มีความแตกต่างกันมากในกระบวนการทำงานที่เกี่ยวกับข้อมูล หากพบว่าการอ่านข้อมูลจากฐานข้อมูลมาประมวลผลมีช่วงเวลาโต้ตอบ (response time) ช้ามากและต้องการปรับการเก็บโดยการจัดทำดัชนีเพิ่มเติมให้การค้นคืนข้อมูลได้สะดวกขึ้นก็จะส่งผลให้การบันทึกข้อมูลกลับต้องช้าลงเพราะต้องเสียเวลาดำเนินการกับดัชนีมากขึ้นกว่าระบบเดิม ด้วยเหตุนี้จึงเป็นเรื่องยากที่เราจะปรับระบบทั้งสองให้มีสมรรถนะดีมากขึ้นพร้อมกัน

ปัจจุบันนี้แนวทางแก้ไขปัญหาข้างต้นก็คือการแยกระบบ OLTP และระบบ OLAP ออกจากกันให้เป็นคนละระบบ โดยให้ระบบ OLTP สามารถจัดเก็บข้อมูลจำนวนมากได้อย่างรวดเร็วมีประสิทธิภาพ และระบบ OLAP ก็สามารถค้นคืนและวิเคราะห์ข้อมูลตามความต้องการของผู้ใช้ได้อย่างรวดเร็ว ระบบ OLTP นั้นปกติยังคงปล่อยให้เป็นแบบเดิม หากใช้คอมพิวเตอร์ขนาดใหญ่เช่นเครื่องเมนเฟรมและใช้ระบบจัดการฐานข้อมูลขนาดใหญ่อยู่แล้วก็เพียงแต่ปรับให้สามารถบันทึกจัดเก็บข้อมูลให้เร็วขึ้น จากนั้นก็จัดทำระบบขึ้นใหม่ให้แยกข้อมูลพื้นฐานออกจากฐานข้อมูลในระบบเดิมแล้วนำข้อมูลมาจัดทำดัชนีใหม่เพื่อให้ผู้บริหารวิเคราะห์ อย่างไรก็ตามทั้งระบบ OLTP และระบบ OLAP ก็อาจจะยังไม่เหมาะที่เราจะนำมาใช้ในการวิเคราะห์ทางธุรกิจหรือช่วยผู้บริหารสำหรับการตัดสินใจ (Decision Support System) ทางธุรกิจได้เพราะต้องใช้เวลาในการประมวลผลที่นานพอสมควรและส่งผลกระทบไปถึงระบบการทำงานของเครื่องที่ใช้งานประจำวัน

เราจะมีวิธีการอย่างไรเพื่อที่จะทำให้ข้อมูลที่เรามีอยู่สามารถนำมาใช้ตอบสนองความต้องการทางธุรกิจได้อย่างรวดเร็วและมีประสิทธิภาพ ดังนั้นจึงได้นำเอาแนวความคิดระบบ ***คลังข้อมูล (data warehouse)*** มาใช้ร่วมกันเพื่อตอบสนองงานในรูปแบบของคลังเก็บข้อมูลสำหรับการบริหารและหากองค์กรใดสามารถที่จะนำข้อมูลที่มีอยู่มาใช้อย่างมีประสิทธิภาพย่อมที่จะทำให้องค์กรประสบความสำเร็จเหนือคู่แข่ง

ข้อมูลส่วนมากที่จัดเก็บในคลังข้อมูลนั้นปกติจะมีน้อยกว่าข้อมูลในฐานข้อมูลของระบบ OLTP เพราะเป็นข้อมูลที่ได้นำมาจัดกลุ่มให้เหมาะสมแก่การค้นคืนแล้ว ข้อมูลเหล่านี้จะมีลักษณะ consistent กล่าวคือ ข้อมูลทุกรายการที่แสดงเรื่องเดียวกันจะต้องเขียนให้เหมือนกัน สะกดแบบเดียวกัน หรือ มีรหัสเดียวกัน หากข้อมูลมีลักษณะแตกต่างกันแล้วจะวิเคราะห์ข้อมูลได้ยาก หรืออาจทำให้ได้ผลลัพธ์ที่ไม่ถูกต้อง ในหน่วยงานและบริษัทขนาดใหญ่นั้นโอกาสที่ข้อมูลทั้งหมดจะ “สะอาด” นั้นเป็นเรื่องที่ยาก ดังนั้นจึงจำเป็นจะต้องมีผู้ทำหน้าที่กลั่นกรองและควบคุมคุณภาพของข้อมูลด้วย

เราสามารถสรุปความแตกต่างของคลังข้อมูลกับฐานข้อมูลที่ใช้ประจำวันได้แต่ละหัวข้อดังนี้

1. Consistency ทั้ง OLTP และ คลังข้อมูล ต่างก็ให้ความสำคัญในเรื่องข้อมูลควรจะมีความสอดคล้องกัน สำหรับ OLTP ซึ่งมีการทำ transaction จำนวนมากๆสิ่งที่ต้องการคือการทำ transaction ให้ครบ ไม่มีการสูญหาย ดังนั้นจึงมีความจำเป็นผู้ส่งและผู้รับจะต้องรับรู้และตรวจสอบอยู่ตลอกเวลาว่าขณะนี้มีการทำ transaction เกิดขึ้นหรือไม่ สำหรับคลังข้อมูล จะไม่สนใจทำการทำ transaction แต่ละครั้ง แต่จะสนใจว่าการ load data ใหม่เข้ามานั้นทำสำเร็จหรือยัง และการ load data เข้ามาทั้งหมดนั้นถูกต้องหรือไม่
2. Transaction สำหรับระบบ OLTP นั้น ในแต่ละวันอาจมีการทำ transaction มากมายซึ่งการทำ transaction แต่ละครั้งจะใช้ข้อมูลเพียงแค่เล็กน้อยเท่านั้น สำหรับคลังข้อมูล แต่ละวันจะทำแค่เพียง 1 transaction ซึ่ง transaction นี้อาจต้องใช้ข้อมูลเป็นจำนวนมากมาย
3. Time Dimension สำหรับ OLTP นั้นจะทำงานอย่างรวดเร็วและทำ transaction อย่างสม่ำเสมอ สถานะของข้อมูลต่างๆมีการเปลี่ยนแปลงอยู่ตลอดเวลา และความสัมพันธ์ระหว่างเอนติตี้ต่างๆก็เปลี่ยนแปลงไปด้วย สำหรับระบบคลังข้อมูลมักจะเก็บข้อมูลในอดีตเพื่อใช้ในการวิเคราะห์ ดังนั้นข้อมูลจะไม่ค่อยมีการเปลี่ยนแปลงตลอดวัน

เป็นที่น่าสังเกตว่าคลังข้อมูลไม่ต้องทำการ normalization เหมือนกับฐานข้อมูลประจำวันที่ต้องทำการ normalization ทั้งนี้เพราะในฐานข้อมูลประจำวัน ข้อมูลจำนวนมหาศาลมีการเปลี่ยนแปลงทำให้ทันสมัยอยู่ตลอดเวลา ดังนั้นประเด็นสำคัญจึงอยู่ที่การเปลี่ยนแปลงทำให้ทันสมัย การออกแบบฐานข้อมูลประจำวันจึงต้องทำให้มีความซ้ำซ้อนหรือ redundancy น้อยที่สุด วิธีการที่จะทำให้ได้ตามจุดประสงค์คือการทำ normalization แต่สำหรับข้อมูลในคลังข้อมูลเป็นข้อมูลที่มีการกลั่นกรองมาแล้ว ใช้ในการวิเคราะห์ตอบคำถามของผู้บริหาร ประเด็นสำคัญจึงไม่อยู่ที่การทำให้ทันสมัย ทำให้ข้อมูลในคลังข้อมูลสามารถมีความซ้ำซ้อนได้ เพราะความซ้ำซ้อนมีข้อดีคือ การตอบคำถามและการออกรายงานสามารถทำได้รวดเร็ว เนื่องจากไม่ต้อง join หลายตาราง ดังนั้นในคลังข้อมูลจึงไม่มีความจำเป็นต้องทำการ normalization

ที่มา http://www.srisangworn.go.th/home/databaselearnx/ms2t1-11.htm