

แผนการสอนประจำบทเรียน

รายชื่ออาจารย์ผู้จัดทำ สุณีย์ พงษ์พินิจภิญโญ

รายละเอียดของเนื้อหา

ตอนที่ 15.1 แนวคิดเกี่ยวกับคลังข้อมูล

เรื่องที่ 15.1.1 ระบบคลังข้อมูลคืออะไร

เรื่องที่ 15.1.2 ประโยชน์ของระบบคลังข้อมูล

ตอนที่ 15.2 สถาปัตยกรรมและองค์ประกอบของคลังข้อมูล

เรื่องที่ 15.2.1 ความสัมพันธ์ระบบคลังข้อมูลกับระบบฐานข้อมูล

เรื่องที่ 15.2.2 โครงสร้างสถาปัตยกรรมและองค์ประกอบของคลังข้อมูล

เรื่องที่ 15.2.3 กระบวนการรวบรวมข้อมูลและวิเคราะห์หารูปแบบข้อมูลในคลังข้อมูล

แนวคิด

1. แนวโน้มในการพัฒนาเทคโนโลยีด้านระบบฐานข้อมูลเพื่อรวบรวมข้อมูลที่มีอยู่ไปช่วยในการตัดสินใจ หรือที่เรียกว่าระบบคลังข้อมูลซึ่งเป็นกระบวนการรวบรวมข้อมูลที่น่าสนใจจากหลากหลายระบบเพื่อนำมาวิเคราะห์หารูปแบบข้อมูลเพื่อช่วยในการตัดสินใจ
2. ข้อมูลที่เอามาใช้ในการตัดสินใจจะถูกรวบรวมไว้ซึ่งอาจอยู่ในรูปที่มีความแตกต่างกันเช่นในรูปแบบของแฟ้มข้อมูล หรือฐานข้อมูล ดังนั้นจึงจำเป็นต้องนำข้อมูลเหล่านี้มาผ่านกระบวนการการแปลงข้อมูลให้อยู่ในมาตรฐานเดียวกัน

วัตถุประสงค์

หลังจากศึกษาบทเรียนที่15 แล้ว นักศึกษาสามารถ

1. บอกถึงลักษณะสำคัญและแนวความคิดของการพัฒนาเทคโนโลยีด้านระบบคลังข้อมูล
2. บอกลักษณะโครงสร้างและสถาปัตยกรรมของคลังข้อมูลได้
3. สามารถออกแบบฐานข้อมูลเพื่อรวบรวมเอาข้อมูลที่น่าสนใจจากหลากหลายระบบเพื่อนำมาวิเคราะห์หารูปแบบข้อมูลเพื่อช่วยในการตัดสินใจ

กิจกรรมการเรียนการสอน

กิจกรรมที่นักศึกษาต้องทำสำหรับการเรียนการสอน ได้แก่

1. ศึกษาเอกสารการสอน
2. ทำกิจกรรมในแบบฝึกปฏิบัติบทเรียนที่ 15
3. ทำแบบประเมินผลของบทเรียนที่ 15

สื่อการสอน

1. เอกสารการสอนของชุดวิชา
2. แบบฝึกปฏิบัติ
3. บทความ/ข้อมูลทางคอมพิวเตอร์
4. การให้คำปรึกษาทางโทรศัพท์
5. CD-ROM
6. Homepage ของชุดวิชาผ่านทางอินเทอร์เน็ต

เอกสารประกอบการสอน

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

ประเมินผล

1. ประเมินผลจากกิจกรรมที่ทำ
2. ประเมินผลจากคำถามท้ายบทเรียนและการสอนประจำภาคการศึกษา

ตอนที่ 15.1 แนวคิดเกี่ยวกับคลังข้อมูล

หัวเรื่อง

- เรื่องที่ 15.1.1 ระบบคลังข้อมูลคืออะไร
เรื่องที่ 15.1.2 ประโยชน์ของระบบคลังข้อมูล

แนวคิด

1. แนวโน้มในการพัฒนาเทคโนโลยีด้านระบบฐานข้อมูลเพื่อรวบรวมข้อมูลที่มีอยู่ไปช่วยในการตัดสินใจ หรือที่เรียกว่าระบบคลังข้อมูลซึ่งเป็นกระบวนการรวบรวมข้อมูลที่สนใจจากหลากหลายระบบเพื่อนำมาวิเคราะห์หารูปแบบข้อมูลเพื่อช่วยในการตัดสินใจ ระบบคลังข้อมูลไม่ได้เป็นโปรแกรมใช้งานสำเร็จรูปที่หลังจากติดตั้งแล้วใช้งานได้เลย แต่เป็นระบบที่สร้างขึ้นตามความต้องการในการวิเคราะห์ของผู้ใช้งานที่หลากหลาย
2. ระบบคลังข้อมูลมีประโยชน์ได้ในหลายหลากธุรกิจ เช่นธุรกิจโทรคมนาคม สามารถใช้วิเคราะห์พฤติกรรมการใช้โทรศัพท์ของผู้ใช้เพื่อจัดกิจกรรมส่งเสริมการตลาดให้เหมาะสม

วัตถุประสงค์

หลังจากที่ศึกษาตอนที่ 15.1 แล้วนักศึกษาสามารถ

1. อธิบายความหมายและแนวความคิดตลอดจนคุณลักษณะของคลังข้อมูลได้
2. บอกถึงความสำคัญและประโยชน์ของคลังข้อมูลที่มีต่อองค์กรนั้นๆ ได้

เรื่องที่ 15.1.1 ระบบคลังข้อมูลคืออะไร

ปัจจุบันนี้องค์กรจะประสบความสำเร็จต้องขึ้นอยู่กับปัจจัยหลายอย่าง และปัจจัยอย่างหนึ่งที่ทำให้องค์กรประสบความสำเร็จคือข้อมูลที่มีอยู่และใช้ประจำวันหรือ Operational Database ซึ่งนับวันจะมีแต่มากขึ้นจนเป็นปัญหาสำหรับองค์กรที่จะต้องจัดการเก็บข้อมูล ซึ่งข้อมูลเหล่านี้อาจเก็บรวบรวมอยู่ในรูปแบบต่างๆไม่ว่าจะเป็นเทปแม่เหล็ก ดิสก์เก็ต หรือในแผ่นซีดี นอกจากนี้ข้อมูลมากมายเหล่านี้ยังไม่เหมาะที่เราจะนำมาใช้ในการวิเคราะห์ทางธุรกิจหรือช่วยผู้บริหารสำหรับการตัดสินใจ (Decision Support System) ทางธุรกิจได้ เพราะต้องใช้เวลาในการประมวลผลที่นานพอสมควรและส่งผลกระทบไปถึงระบบการทำงาน นกข เครื่องที่ใช้งานประจำวันอีกด้วย

เราจะมีวิธีอย่างไรเพื่อที่จะทำให้ข้อมูลที่เราที่มีอยู่สามารถนำมา ใช้ตอบสนองความต้องการทางธุรกิจได้อย่างรวดเร็วและมีประสิทธิภาพ ดังนั้นแนวความคิดเรื่องคลังข้อมูล (Data Warehouse) จึงเกิดขึ้นเพื่อตอบสนองงานในรูปแบบของคลังเก็บข้อมูลสำหรับการบริหารและหากองค์กรใดสามารถที่จะนำข้อมูลที่มีอยู่มาใช้ อย่างมีประสิทธิภาพย่อมที่จะทำให้องค์กรประสบความสำเร็จเหนือคู่แข่ง

อย่างไรก็ตามเมื่อเรามีข้อมูลแล้วแต่ถ้าเราไม่มีการจัดการหรือการบริหารข้อมูลที่ดี ก็จะทำให้องค์กรยากต่อการดำเนินธุรกิจ สำหรับในประเทศไทยเองแนวความคิดคลังข้อมูลได้ถูกนำเข้ามาใช้กับหน่วยงานขนาด

ใหญ่เช่นธนาคาร บริษัท เงินทุนหลักทรัพย์หรือหน่วยงานที่จำเป็นต้องอาศัยข้อมูลทำการวิเคราะห์ยอดขายในอนาคต การเปรียบเทียบยอดขายในปัจจุบันกับยอดขายที่เกิดขึ้นในอดีต หรือการเปรียบเทียบข้อมูลในส่วนต่างๆ ของบริษัทแล้วนำมาวิเคราะห์ในรูปแบบข้อมูลสำหรับการบริหารเพื่อเป็นแนวทางสำหรับการกำหนดแผนงานต่อไปในอนาคต

คลังข้อมูลคืออะไร

คลังข้อมูลคือที่เก็บข้อมูลขององค์กรที่ได้รับการออกแบบเพื่อช่วยการตัดสินใจของฝ่ายบริหารในทางปฏิบัติ นั่นคือ สิ่งที่เก็บอยู่ในคลังข้อมูลไม่ได้มีแต่เพียงข้อมูลเท่านั้น หากยังเก็บเครื่องมือสำหรับดำเนินการกับข้อมูล กระบวนการทำงานกับข้อมูล และทรัพยากรอื่นๆ ระบบคลังข้อมูลเพื่อการบริหารได้ถูกออกแบบมาเพื่อใช้ในการเก็บข้อมูลขนาดใหญ่ โดยระบบข้อมูลเพื่อการบริหารนี้จะแยกข้อมูลออกจากฐานข้อมูลที่ใช้งานประจำวัน (Operational Database) ซึ่งข้อมูลสำหรับการบริหารโดยมากจะเป็นข้อมูลสรุป (Summary Data) ข้อมูลสรุปนี้อาจจะเป็นข้อมูลในอดีต ข้อมูลอ้างอิง หรือข้อมูล ณ ปัจจุบัน ซึ่งอาจได้มาจากข้อมูล Operational Database หรือมีการประมวลผลข้อมูลใน Operational Database ให้เป็นข้อมูลสรุป หรืออาจนำมาจากที่อื่นภายนอกองค์กรและทำการเพิ่มเติมลงไปก็ได้ ข้อมูลที่ถูกจัดเก็บอยู่ในคลังข้อมูลถือว่าเป็นข้อมูลในรูปแบบ Relational Database Management Systems (RDBMS) ที่มีประสิทธิภาพสูง โดยมากเราจะเลือกเก็บแต่เฉพาะข้อมูลที่สำคัญสำหรับการตัดสินใจหรือหัวข้อของธุรกิจที่น่าสนใจ

วัตถุประสงค์ของการสร้างคลังข้อมูล

เป้าหมายของการสร้างคลังข้อมูลคือ การแยกกลุ่มข้อมูลสารสนเทศที่ใช้ในการวิเคราะห์ทางธุรกิจออกจากฐานข้อมูลที่ใช้งานประจำวัน (Operational Database) มาเก็บอยู่ใน Relational Database Management Systems (RDBMS) ประสิทธิภาพสูง และทำให้การเรียกใช้ข้อมูลชุดนี้ทำได้ง่ายและรวดเร็ว จากเครื่องมือที่อยู่บนเครื่องเดสก์ท็อปทั่วไป โดยลด off-loading เพิ่มกลไกการช่วยตัดสินใจ ปรับปรุงเวลาที่ตอบสนอง (response time) รวดเร็วยิ่งขึ้นอย่างมากและผู้บริหารสามารถเรียกข้อมูลรายละเอียดที่จำเป็น ที่ถูกเก็บมาก่อนหน้านี้ (historical data) มาใช้ช่วยในการตัดสินใจทางธุรกิจแม่นยำขึ้น

เป้าหมายในการสร้างคลังข้อมูลมีดังนี้

1. คลังข้อมูลทำให้สามารถเข้าถึงข้อมูลขององค์กรได้ ผู้จัดการและนักวิเคราะห์ขององค์กรสามารถเชื่อมต่อเข้าไปยังคลังข้อมูลจากเครื่องคอมพิวเตอร์ของตนได้ ซึ่งการเชื่อมต่อสามารถทำได้ทันทีตามความต้องการและด้วยประสิทธิภาพสูง เครื่องมือที่มีให้กับผู้จัดการและนักวิเคราะห์ใช้งานง่าย สามารถออกรายงานได้ด้วยการคลิกปุ่มเดียว
2. ข้อมูลในคลังข้อมูลมีความถูกต้องตรงกันหมด คำถามเดียวกันต้องได้รับคำตอบที่เหมือนกันเสมอไม่ว่าผู้ถามจะเป็นใคร ถามเวลาใด
3. ข้อมูลในคลังข้อมูลสามารถถูกวิเคราะห์จากหัวข้อในธุรกิจประเภทนั้น โดยแบ่งข้อมูลหรือรวมข้อมูลมาวิเคราะห์ตามความต้องการ

4. คลังข้อมูลเป็นส่วนที่ผลิตข้อมูลจาก OLTP ข้อมูลไม่เพียงแต่ถูกรวบรวมมาไว้ที่ศูนย์กลางอย่างเดียว แต่จะถูกรวบรวมอย่างระมัดระวังจากแหล่งข้อมูลหลายๆแห่งนอกองค์กรด้วย แล้วมาปรับปรุงให้เหมาะสมกับการใช้งานเท่านั้น ถ้าข้อมูลเชื่อถือไม่ได้หรือไม่สมบูรณ์จะไม่ถูกอนุญาตให้นำไปใช้
5. คุณภาพของข้อมูลในคลังข้อมูลเป็นตัวผลักดันให้สามารถทำการ reengineering ธุรกิจได้

เรื่องที่ 15.1.2 ประโยชน์ของระบบคลังข้อมูล

โดยทั่วไปแล้วข้อมูล Operational Database จะเก็บข้อมูลในรูปแบบ Transaction Systems เมื่อมีความต้องการข้อมูลในอันที่จะนำมาใช้ช่วยในการตัดสินใจก็จะประสบปัญหาต่างๆเช่น

- บุคลากรทางด้าน Information Systems จำเป็นต้องเรียกข้อมูลจากฐานข้อมูลขนาดใหญ่ ซึ่งมีข้อมูลมากเกินความต้องการ ส่งผลให้ประสิทธิภาพของ Transaction Operational Database ทำงานได้ช้าลง
- ข้อมูลจะเป็นรูปแบบข้อมูลตารางเท่านั้น
- ข้อมูลจะถูกนำเสนอในรูปแบบที่ตายตัว ไม่สามารถเปลี่ยนแปลงตามความต้องการของผู้ใช้
- ไม่ตอบสนองความต้องการของการตัดสินใจ เพราะข้อมูลสำหรับการตัดสินใจมีความสลับซับซ้อนสูง มีการรวมตัวกันของข้อมูลจากตารางต่างๆหลายๆตารางข้อมูล
- ไม่ตอบสนองการสอบถามข้อมูล (Data Queries) สำหรับผู้ใช้
- มีข้อมูลย้อนหลังน้อย (Historical Data)
- ข้อมูลถูกจัดเก็บกระจายตามที่ต้องการ ซึ่งยากต่อการเรียกใช้หรือขาดความสัมพันธ์ทางธุรกิจ อันอาจจะต้องเสียเวลาในการทำให้สอดคล้อง หรือเกิดความซ้ำซ้อนของข้อมูลได้

จากอุปสรรคที่กล่าวมาข้างต้นคลังข้อมูลจึงได้ถูกออกแบบมาเพื่อตอบสนองงานในรูปแบบการตัดสินใจโดยการแยกฐานข้อมูลออกจาก Operational Database และเก็บข้อมูลในรูปแบบข้อมูลสรุป (Summary Data) ซึ่งข้อมูลสรุปนี้จะเลือกแต่เฉพาะข้อมูลที่มีความสำคัญต่อการตัดสินใจหรือเพื่อใช้ในการบริหารไปจนถึงการกำหนดแผนงานในอนาคต

ในระบบคลังข้อมูล ข้อมูลที่ซับซ้อนจะถูกรวบรวมหรือเปลี่ยนแปลงให้ง่ายต่อการจัดเก็บและสามารถเรียกกลับมาใช้ได้อย่างรวดเร็วและถูกต้อง โดยข้อมูลต่างๆเหล่านี้จะถูกนำมาใช้สำหรับการวิเคราะห์และช่วยในเรื่องการตัดสินใจโดยอาศัยเครื่องมือ (tool) ที่อยู่ในเครื่องคอมพิวเตอร์ที่เป็นซอฟต์แวร์มาใช้ในการจัดการทำรายงานและเพิ่มประสิทธิภาพสำหรับการตัดสินใจให้รวดเร็วยิ่งขึ้น โดยผู้บริหาร นักวางแผน และนักวิเคราะห์ข้อมูลสามารถเรียกหาข้อมูลหรือสอบถาม (query) เพื่อให้ได้รับคำตอบในรูปแบบตารางรายงาน หรือรายงาน กราฟ เพื่อมาทำการวิเคราะห์ข้อมูลด้วยตนเองเช่น

- การเปรียบเทียบยอดขายระหว่างช่วงเวลาในอดีตกับปัจจุบันไปจนถึงการทำพยากรณ์ยอดขายในอดีต (Forecasting)
- การหายอดขายสูงสุดหรือต่ำสุด
- การเปรียบเทียบยอดขาย ต้นทุน กำไร ในรูปแบบตารางรายงาน หรือรายงาน กราฟ

ซึ่งเครื่องมือนี้ถือได้ว่าเป็นสิ่งสำคัญในอันที่จะนำองค์กรไปสู่ความสำเร็จในกระบวนการตัดสินใจ ในปัจจุบันเครื่องมือที่ตอบสนองงานเพื่อช่วยผู้บริหารสำหรับการตัดสินใจมีอยู่มากมายในตลาด ทั้งนี้ก็เป็นทางเลือกของผู้ใช้ในการที่จะเลือกเครื่องมือที่มีประสิทธิภาพสูงสุดเพื่อตอบสนองงานของผู้บริหารในกระบวนการตัดสินใจต่อไป

จะเห็นได้ว่าการจัดทำคลังข้อมูลเป็นความท้าทายอย่างหนึ่งของหน่วยงาน ทั้งนี้เพราะหน่วยงานต่าง ๆ มักจะมีข้อมูลธุรกรรมที่ไม่มีความต่อเนื่องกัน (consistent) และมีความลึกลับอยู่มากดังได้อธิบายไปข้างแล้ว ดังนั้นการจัดทำคลังข้อมูลจะต้องหาทางแก้ปัญหาให้ได้ อีกประการหนึ่งก็คือข้อมูลบางส่วนหายไปหรือมีไม่ครบ ยกตัวอย่างบริษัทแห่งหนึ่งต้องการวิเคราะห์ความสนใจของลูกค้าที่ใช้บัตรเครดิตของบริษัทออกให้ โดยกำหนดจะแยกความสนใจว่ามีความแตกต่างระหว่างเพศหรือไม่ แต่ในการจัดทำระบบประมวลผลธุรกรรมตั้งแต่แรกนั้นนักวิเคราะห์ระบบไม่ได้กำหนดให้เก็บข้อมูลเพศของลูกค้าเอาไว้เพราะเห็นว่าไม่เกี่ยวกับธุรกรรม ดังนั้นผู้จึงไม่สามารถนำข้อมูลมาวิเคราะห์ได้ ในกรณีเช่นนี้ระหว่างการจัดทำคลังข้อมูลก็ต้องจัดให้มีพนักงานที่ทำหน้าที่ศึกษาข้อมูลโดยพิจารณาจากแบบฟอร์มเดิมแล้วนำเพศมาบันทึกเป็นข้อมูลเพิ่มเติมขึ้น

การจัดทำคลังข้อมูลจะมีความสำคัญมากขึ้นในอนาคต เพราะปัจจุบันนี้ผู้ใช้และผู้บริหารของหน่วยงานเริ่มมีเข้าใจความสำคัญของข้อมูลมากขึ้น และเริ่มตระหนักว่าหากนำข้อมูลมาวิเคราะห์ให้เข้าใจสถานการณ์หรือเหตุการณ์ที่เกิดขึ้นแล้วจะทำให้หน่วยงานหรือบริษัทสามารถตอบสนองต่อเหตุการณ์ที่เกิดขึ้นได้ดียิ่งขึ้น และจะทำให้หน่วยงานหรือบริษัททำงานบรรลุวัตถุประสงค์และเป้าหมายได้ดียิ่งขึ้นตามไปด้วย

กิจกรรม 15.1

1. คลังข้อมูลคืออะไร
2. คลังข้อมูลแตกต่างจากฐานข้อมูล (Operational Database) ที่ใช้งานประจำวันอย่างไร

แนวตอบกิจกรรม 15.1

1. คลังข้อมูลคือที่เก็บข้อมูลขององค์กร ที่ได้รับการออกแบบเพื่อช่วยการตัดสินใจของฝ่ายบริหารในทางปฏิบัตินั้น สิ่งทีเก็บอยู่ในคลังข้อมูลไม่ได้มีแต่เพียงข้อมูลเท่านั้น หากยังเก็บเครื่องมือสำหรับดำเนินการกับข้อมูล กระบวนการทำงานกับข้อมูล และทรัพยากรอื่นๆ
2. คลังข้อมูลได้ถูกออกแบบมาเพื่อตอบสนองงานในรูปแบบการตัดสินใจโดยการแยกฐานข้อมูลออกจาก Operational Database และเก็บข้อมูลในรูปแบบข้อมูลสรุป (Summary Data) ซึ่งข้อมูลสรุปนี้จะเลือกแต่เฉพาะข้อมูลที่มีความสำคัญต่อการตัดสินใจหรือเพื่อใช้ในการบริหารไปจนถึง

การกำหนดแผนงานในอนาคต ส่วนข้อมูล Operational Database จะเก็บข้อมูลในรูปแบบ Transaction Systems ซึ่งเป็นฐานข้อมูลที่ใช้กับงานประจำวัน

ตอนที่ 15.2 สถาปัตยกรรมและองค์ประกอบของคลังข้อมูล

หัวเรื่อง

- เรื่องที่ 15.2.1 ความสัมพันธ์ระบบคลังข้อมูลกับระบบฐานข้อมูล
- เรื่องที่ 15.2.2 โครงสร้างสถาปัตยกรรมและองค์ประกอบของคลังข้อมูล
- เรื่องที่ 15.2.3 กระบวนการรวบรวมข้อมูลและวิเคราะห์รูปแบบข้อมูลในคลังข้อมูล

แนวคิด

1. สถาปัตยกรรมและองค์ประกอบของคลังข้อมูล ข้อมูลที่เอามาใช้ในการตัดสินใจจะถูกรวบรวมไว้ซึ่งอาจอยู่ในรูปที่มีความแตกต่างกันเช่นในรูปแบบของแฟ้มข้อมูล หรือฐานข้อมูล ดังนั้นจึงจำเป็นต้องนำข้อมูลเหล่านี้มาผ่านกระบวนการการแปลงข้อมูลให้อยู่ในมาตรฐานเดียวกัน เทคโนโลยีในระบบคลังข้อมูลจะประกอบด้วยสามส่วนหลักคือ Data Extraction Tool ระบบฐานข้อมูล และ Analysis Tool โดยที่ Data Extraction Tool ใช้ดึงข้อมูลที่น่าสนใจเพื่อการวิเคราะห์จากต้นทางและทำการแปลงข้อมูลก่อนเก็บข้อมูลไว้ในระบบฐานข้อมูลเพื่อใช้ Analysis Tool มาวิเคราะห์ข้อมูลอีกทีหนึ่ง ดังนั้นระบบฐานข้อมูลจึงเป็นแหล่งเก็บข้อมูลที่สนใจและคัดเลือกแล้วเพื่อใช้วิเคราะห์ในระบบคลังข้อมูล

วัตถุประสงค์

หลังจากศึกษาตอนที่ 15.2 แล้ว นักศึกษาสามารถ

1. บอกถึงรูปแบบของความสัมพันธ์ระหว่างระบบคลังข้อมูลกับระบบฐานข้อมูลได้
2. บอกโครงสร้างสถาปัตยกรรมและองค์ประกอบรูปแบบของคลังข้อมูล
3. เข้าใจกระบวนการนำข้อมูลจากฐานข้อมูลมาแปรรูปและวิเคราะห์ให้เป็นรูปแบบในคลังข้อมูล
4. ออกแบบและพัฒนาคลังข้อมูลได้

เรื่องที่ 15.2.1 ความสัมพันธ์ระบบคลังข้อมูลกับระบบฐานข้อมูล

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

เนื่องจากฐานข้อมูลทั่วไป (Operational Database) มีลักษณะดังได้กล่าวมาแล้วจึงมีความสามารถเพียงแค่การเรียกใช้ข้อมูลที่มีอยู่ แต่ไม่สามารถจะนำมาช่วยในการสนับสนุนการตัดสินใจได้ เพราะเมื่อมีการเรียกใช้ข้อมูลจะต้องเรียกใช้ข้อมูลจากฐานข้อมูลขนาดใหญ่ ซึ่งมีข้อมูลจำนวนมากและมีการแตกตารางที่

นอร์มัลไลซ์ (normalized table) แล้วออกเป็นหลายตาราง จึงไม่รองรับคำถามที่ต้องการจะนำมาช่วยในการสนับสนุนการตัดสินใจ (decision support queries) มีการรวม (join) กันของตารางต่างๆที่ซับซ้อน ซึ่งจะทำให้มีประสิทธิภาพของการค้นหาข้อมูลจากฐานข้อมูลน้อยลง และทำงานช้าลง ไม่สามารถเรียกใช้ข้อมูลที่ต้องการได้ทั้งหมดเพราะมีรูทีนอัตโนมัติ (Automate Routine) จึงมีความสามารถในการค้นหาข้อมูลแบบที่ไม่ซ้ำซ้อนเท่านั้น นอกจากนี้การเก็บข้อมูลในระบบฐานข้อมูลทั่วไป (Operational Database) ยังไม่มีการเก็บข้อมูลย้อนหลัง (historical data) เพื่อช่วยในการคาดคะเนแนวโน้มที่คาดว่าจะเป็นไปได้ในอนาคต

ดังนั้นระบบคลังข้อมูลจึงได้ถูกคิดขึ้นมาเพื่อช่วยให้ผู้ใช้เรียกใช้ข้อมูลที่มีอยู่ได้อย่างมีประสิทธิภาพสูงสุดด้วยวิธีที่สร้างสรรค์เพราะธรรมชาติที่แตกต่างกันระหว่างระบบฐานข้อมูลคลังข้อมูลและระบบฐานข้อมูลทั่วไป ดังนั้นฐานข้อมูลคลังข้อมูลจะต้องมีคุณสมบัติดังนี้

1. Subject Oriented ข้อมูลจะต้องถูกสร้างขึ้นจากหัวข้อ (subject) ธุรกิจที่สนใจ เช่น ถ้าบริษัทประกันภัยต้องการใช้คลังข้อมูล ฐานข้อมูลที่ได้จะต้องสร้างขึ้นจากประวัติลูกค้า, เบี้ยประกัน และการเรียกร้องแทนที่จะแยกตามชนิดของผลิตภัณฑ์ หรือบริการประกันภัย/ประกันชีวิต ข้อมูลที่สร้างขึ้นจะประกอบด้วยหัวข้อที่เก็บเฉพาะข่าวสารที่จำเป็น สำหรับกระบวนการตัดสินใจเท่านั้น
2. Integrated ข้อมูลถูกรวบรวมจากแหล่งต่างๆ จากระบบปฏิบัติการ, รูปแบบของข้อมูล, แพลตฟอร์มที่หลากหลาย สร้างขึ้นเป็นฐานข้อมูลที่สุดคล้องเป็นหนึ่งเดียว เช่นค่าของตัวแปรตัวเดียวในแต่ละฐานข้อมูลอาจต่างกัน ฐานข้อมูลหนึ่งอาจใช้ 0 และ 1 อีกฐานข้อมูลหนึ่งอาจใช้ T และ F ดังนั้นฐานข้อมูลที่จะสร้างใหม่จะต้องได้รับการกำหนดค่าตัวแปรให้เหมือนกันเป็นหนึ่งเดียว
3. Time-variant ข้อมูลซึ่งใช้ตัดสินใจที่เก็บไว้จะต้องมีอายุประมาณ 5 ถึง 10 ปี เพื่อให้เปรียบเทียบหาแนวโน้ม และทำนายผลลัพธ์ในอนาคตได้
4. Non-volatile ข้อมูลจะไม่อัปเดตหรือถูกทำให้เปลี่ยนแปลงง่ายๆ ผู้ใช้สามารถใช้ฐานข้อมูลคลังข้อมูลได้เพียงแค่อ่านและเข้าถึงเท่านั้น

โดยระบบฐานข้อมูลคลังข้อมูลจะแยกกลุ่มข้อมูลสารสนเทศที่ใช้ในการวิเคราะห์ทางธุรกิจออกจากฐานข้อมูลที่ใช้ประจำวัน (Operational Database) มาเก็บอยู่ในระบบจัดการฐานข้อมูล (Relational Database Management Systems) ประสิทธิภาพสูงสุด และทำให้การเรียกใช้ข้อมูลชุดนี้ทำได้ง่ายยิ่งขึ้น จากเครื่องมือที่อยู่บนเครื่องคอมพิวเตอร์เดสก์ท็อปทั่วไป โดยลด off-loading เพิ่มกลไกการช่วยตัดสินใจ ปรับปรุงเวลาที่ตอบสนอง (response time) รวดเร็วขึ้นอย่างมากและผู้บริหารสามารถเรียกข้อมูลรายละเอียดที่จำเป็นที่ถูกเก็บมาก่อนหน้านี้ (historical data) มาช่วยในการตัดสินใจทางธุรกิจแม่นยำขึ้น

ความแตกต่างอีกประการหนึ่งก็คือผู้ใช้คลังข้อมูลมักจะต้องการจัดกลุ่มข้อมูลด้วยตนเองมากกว่าผู้ใช้ในระบบฐานข้อมูลธรรมดา ยกตัวอย่างผู้ใช้ อาจต้องการวิเคราะห์ผลกระทบของการทำการตลาดแบบต่างๆ อาจต้องการจัดกลุ่มการขายสินค้าแยกตามผลิตภัณฑ์ หรือรูปแบบของการจัดผลิตภัณฑ์ เช่น การห่อรวมสินค้าไว้ในบรรจุภัณฑ์ที่ต่างกัน หรือการรวมผลิตภัณฑ์ต่างรูปแบบไว้ด้วยกัน ในกรณีต่างๆ เหล่านี้ผู้ใช้ต้องการที่จะเลือกจัดกลุ่มข้อมูลได้ตามใจชอบ นอกจากการนำข้อมูลเข้ามารวมกันแล้ว ผู้ใช้ยังอาจต้องการที่จะแยกแยะข้อมูลในแบบ

ที่ตนเองต้องการได้ ยกตัวอย่างในการจัดทำคลังข้อมูลเกี่ยวกับนักวิจัยและผลงานวิจัยของประเทศ หน่วยงานอาจจัดเก็บข้อมูลเอาไว้เป็นกลุ่มก้อนโดยไม่ได้แยกสาขา แต่ต่อมาผู้ใช้อาจต้องการนำข้อมูลนักวิจัยมาวิเคราะห์แยกแยะว่าทั้งประเทศมีนักวิจัยสาขาต่างๆ เป็นจำนวนเท่าใด ทำงานวิจัยด้านใดบ้าง ใช้เงินด้านวิจัยไปเท่าใด เป็นต้น โดยปกติแล้วการจัดทำฐานข้อมูลให้สามารถวิเคราะห์แยกแยะข้อมูลในแบบนี้ได้นั้นเป็นเรื่องไม่ยาก แต่ในการออกแบบคลังข้อมูลนั้นจำเป็นต้องเผื่อให้ผู้ใช้หลายคนสามารถแยกแยะข้อมูลตามความต้องการที่แตกต่างกันได้ด้วย ผู้ใช้จำนวนมากในปัจจุบันนี้อาจใช้ซอฟต์แวร์หลากหลายประเภทสำหรับเครื่องคอมพิวเตอร์ส่วนบุคคล ผู้ใช้บางคนอาจจะใช้โปรแกรมสเปรดชีตในการวิเคราะห์ข้อมูล และผู้ใช้งานบางคนอาจต้องการใช้โปรแกรมวิเคราะห์สถิติอื่นๆ ดังนั้นผู้ใช้นี้จึงอาจมีความต้องการในการนำเข้าสู่ข้อมูลจากคลังข้อมูลมาไว้ในแฟ้มข้อมูลที่มีรูปแบบตรงกับโปรแกรมที่ตนต้องการใช้ ความต้องการด้านนี้นับว่าสำคัญมากที่สุดในการจัดทำคลังข้อมูล

งานอย่างหนึ่งที่นิยมใช้ฐานข้อมูลกันมากก็คืองานบันทึกข้อมูลธุรกรรมเอาไว้เพื่อประมวลผล ข้อมูลธุรกรรมเหล่านี้ได้แก่ ข้อมูลการสั่งซื้อสินค้าของลูกค้า ข้อมูลการซื้อบัตรโดยสารเครื่องบิน ข้อมูลการฝากหรือถอนเงินของลูกค้าธนาคาร แต่เดิมนั้นการบันทึกข้อมูลธุรกรรมเริ่มต้นด้วยการใช้กระดาษแบบฟอร์มสำหรับให้ลูกค้ากรอกข้อมูล จากนั้นจึงนำแบบฟอร์มมาบันทึกข้อมูลลงในฐานข้อมูลของระบบคอมพิวเตอร์ในแบบแบตช์ (batch) ปัจจุบันนี้การบันทึกข้อมูลธุรกรรมได้เปลี่ยนไปเป็นระบบออนไลน์ (online) เป็นส่วนใหญ่ ในระบบแบบนี้กระบวนการบันทึกข้อมูลมีลักษณะอัตโนมัติมากขึ้นและให้อุปกรณ์บันทึกข้อมูลที่สามารถเก็บข้อมูลลงในฐานข้อมูลของระบบคอมพิวเตอร์ได้ทันที เช่น การใช้อุปกรณ์ฝากถอนเงินโดยอัตโนมัติ (ATM) ทำให้สามารถประมวลผลการฝากถอนเงินและบันทึกข้อมูลที่เกิดขึ้นได้ทันที หรือในห้างสรรพสินค้าก็มีการใช้เครื่องบริการ ณ จุดขาย (Point of Sale; POS) สำหรับอ่านรหัสแท่ง แสดงราคาสินค้า แล้วบันทึกข้อมูลการขายไปเก็บไว้ในฐานข้อมูลได้ทันที การดำเนินการในลักษณะนี้เรียกกันว่า การประมวลผลธุรกรรมออนไลน์ (On-Line Transaction Processing; OLTP)

1. ลักษณะงานการประมวลผลธุรกรรมออนไลน์และการประมวลผลเชิงวิเคราะห์ออนไลน์

ระบบ OLTP โดยทั่วไปจะต้องสามารถดำเนินการกับข้อมูลธุรกรรมได้อย่างมีประสิทธิภาพ งานที่ทำงานกับข้อมูลได้แก่การปรับค่าของข้อมูลให้เป็นปัจจุบันและการเพิ่มข้อมูลลงไปในฐานข้อมูล ข้อมูลเหล่านี้จึงจะมีจำนวนมากและเพิ่มขึ้นตลอดเวลา ณ เวลาใดเวลาหนึ่งอาจจะมีประมวลผลข้อมูลจำนวนนับแสนเรคอร์ดได้ เช่น ณ สนามบินแต่ละแห่งจะมีผู้โดยสารเข้ามารับบัตรที่นั่งของสายการบินต่างๆ เป็นจำนวนนับหมื่นๆ คน คอมพิวเตอร์ของสายการบินจะต้องตรวจสอบการสำรองที่นั่ง ต้องบันทึกเลขที่นั่งและเที่ยวบินรวมทั้งจะต้องปรับเปลี่ยนโยกย้ายข้อมูลจากเที่ยวบินหนึ่งไปอีกเที่ยวบินหนึ่งได้ด้วย หรือในกรณีของศูนย์การค้า และซูเปอร์มาร์เก็ต จะมีการบันทึกเรคอร์ดการขายเพิ่มเข้าไปในฐานข้อมูลการขายตลอดเวลา รวมแล้ววันละเป็นหมื่นๆ รายการ การออกแบบระบบ OLTP แบบนี้จำเป็นต้องหาทางให้ระบบสามารถทำงานได้อย่างถูกต้องรวดเร็วตลอดเวลา เพื่ออำนวยความสะดวกให้ผู้ใช้งานจำนวนมากสามารถใช้ระบบได้พร้อมกัน อีกทั้งยังต้องสามารถแก้ไขพื้นที่สภาพให้กลับตั่งเดิมได้หากเกิดความขัดข้องเสียหาย

การที่จะจัดทำระบบ OLTP ให้มีความสามารถในแบบนี้ได้ต้องคำนึงถึงปัจจัยต่อไปนี้

- 1) ขนาดและตำแหน่งของ rollback segment
- 2) ดัชนี การจัดกลุ่ม และ การคำนวณตำแหน่งที่อยู่ (hashing)
- 3) การออกแบบข้อมูลธุรกรรมให้เหมาะกับงานประยุกต์
- 4) หน่วยเก็บและเนื้อที่ว่างสำหรับการเก็บข้อมูลใหม่
- 5) ความเข้าใจลักษณะงานประยุกต์และการเขียนคำสั่งสำหรับค้นคืนข้อมูล
- 6) การปรับปรุงสมรรถนะของระบบอย่างต่อเนื่อง

ระบบ OLTP ที่พัฒนาขึ้นโดยใช้เทคนิคด้านฐานข้อมูลตามปกติมักจะไม่สามารถรับกับปริมาณข้อมูลที่เพิ่มขึ้นอย่างมากมาเป็นประจำทุกวันได้ การนำระบบเช่นนี้มาใช้จึงมีความเสี่ยงที่จะเกิดความผิดพลาดเสียหายขึ้น วิธีการแก้ไขก็คือการแยกฐานข้อมูลออกมาเป็นส่วน ๆ ให้เหมาะกับการใช้งาน

งานที่เกี่ยวข้องกับฐานข้อมูลอีกอย่างหนึ่งก็คืองานที่เรียกว่า *การประมวลผลเชิงวิเคราะห์ออนไลน์ (On-Line Analytical; OLAP)* ระบบ OLTP ที่กล่าวไปแล้วนั้นเน้นที่การบันทึกเก็บข้อมูลใหม่ๆ เพิ่มเข้าไปในฐานข้อมูล ส่วนระบบ OLAP นั้นเน้นที่การค้นคืนข้อมูลที่มีอยู่แล้วจากฐานข้อมูลเพื่อนำมาวิเคราะห์อย่างละเอียด ผู้ใช้ระบบ OLAP ส่วนใหญ่คือผู้บริหาร นักวิจัยตลาด นักสถิติ หรือ ผู้ใช้อื่นๆ ดังนั้นปัจจัยสำคัญสำหรับความสำเร็จของระบบ OLAP ก็คือระบบจะต้องทำงานได้รวดเร็ว สามารถค้นหาข้อมูลจากฐานข้อมูลขนาดใหญ่มาคำนวณได้อย่างครบถ้วนไม่ตกหล่น ในขณะที่เดียวกันระบบก็จะต้องมีความมั่นคง ไม่ผิดพลาดได้ง่ายระหว่างการใช้งาน

ปัจจัยที่จะทำให้ทำได้ตามนี้ก็กล่าวนี้มีอยู่สามข้อคือ

- 1) จะต้องมียระบบจัดคำสั่งค้นคืนข้อมูลให้ทำงานได้รวดเร็วที่สุด (query optimization)
- 2) การจัดดัชนี จัดกลุ่มข้อมูล และ การคำนวณตำแหน่งที่อยู่ข้อมูล
- 3) การประมวลผลคำสั่งค้นคืนในแบบขนาน โดยเฉพาะเมื่อใช้หน่วยเก็บแบบ RAID

แม้ว่าระบบ OLTP และ OLAP นี้จะเกี่ยวข้องกับข้อมูลธุรกรรมเหมือนกันแต่ก็มีความแตกต่างกันมากในกระบวนการทำงานที่เกี่ยวกับข้อมูล หากพบว่าการอ่านข้อมูลจากฐานข้อมูลมาประมวลผลมีช่วงเวลาได้ตอบ (response time) ช้ามากและต้องการปรับการเก็บโดยการจัดทำดัชนีเพิ่มเติมให้การค้นคืนข้อมูลได้สะดวกขึ้น ก็ส่งผลให้การบันทึกข้อมูลกลับต้องช้าลงเพราะต้องเสียเวลาดำเนินการกับดัชนีมากขึ้นกว่าระบบเดิม ด้วยเหตุนี้จึงเป็นเรื่องยากที่เราจะปรับระบบทั้งสองให้มีสมรรถนะดีขึ้นพร้อมกัน

ปัจจุบันนี้แนวทางแก้ไขปัญหานี้ข้างต้นก็คือการแยกระบบ OLTP และระบบ OLAP ออกจากกันให้เป็นคนละระบบ โดยให้ระบบ OLTP สามารถจัดเก็บข้อมูลจำนวนมากได้อย่างรวดเร็วมีประสิทธิภาพ และระบบ OLAP ก็สามารถค้นคืนและวิเคราะห์ข้อมูลตามความต้องการของผู้ใช้ได้อย่างรวดเร็ว ระบบ OLTP นั้นปกติยังคงปล่อยให้ระบบเดิม หากใช้คอมพิวเตอร์ขนาดใหญ่เช่นเครื่องเมนเฟรมและใช้ระบบจัดการฐานข้อมูลขนาดใหญ่อยู่แล้วก็เพียงแค่ปรับให้สามารถบันทึกจัดเก็บข้อมูลให้เร็วขึ้น จากนั้นก็จัดทำระบบขึ้นใหม่ให้แยกข้อมูลพื้นฐานออกจากฐานข้อมูลในระบบเดิมแล้วนำข้อมูลมาจัดทำดัชนีใหม่เพื่อให้ผู้บริหารวิเคราะห์ อย่างไรก็ตามทั้งระบบ OLTP และระบบ OLAP ก็อาจจะยังไม่เหมาะที่เราจะนำมาใช้ในการวิเคราะห์ทางธุรกิจหรือช่วยผู้บริหาร

สำหรับการตัดสินใจ (Decision Support System) ทางธุรกิจได้เพราะต้องใช้เวลาในการประมวลผลที่นานพอสมควรและส่งผลกระทบไปถึงระบบการทำงานของเครื่องที่ใช้งานประจำวัน

เราจะมีวิธีการอย่างไรเพื่อที่จะทำให้ข้อมูลที่เรามีย่อยสามารถนำมาใช้ตอบสนองความต้องการทางธุรกิจได้อย่างรวดเร็วและมีประสิทธิภาพ ดังนั้นจึงได้นำเอาแนวความคิดระบบ **คลังข้อมูล (data warehouse)** มาใช้ร่วมกันเพื่อตอบสนองงานในรูปแบบของคลังเก็บข้อมูลสำหรับการบริหารและหากองค์กรใดสามารถที่จะนำข้อมูลที่มีอยู่มาใช้อย่างมีประสิทธิภาพย่อมจะทำให้องค์กรประสบความสำเร็จเหนือคู่แข่ง

ข้อมูลส่วนมากที่จัดเก็บในคลังข้อมูลนั้นปกติจะมีน้อยกว่าข้อมูลในฐานข้อมูลของระบบ OLTP เพราะเป็นข้อมูลที่ได้นำมาจัดกลุ่มให้เหมาะสมแก่การค้นคืนแล้ว ข้อมูลเหล่านี้จะมีลักษณะ consistent กล่าวคือ ข้อมูลทุกรายการที่แสดงเรื่องเดียวกันจะต้องเขียนให้เหมือนกัน สะกดแบบเดียวกัน หรือ มีรหัสเดียวกัน หากข้อมูลมีลักษณะแตกต่างกันแล้วจะวิเคราะห์ข้อมูลได้ยาก หรืออาจทำให้ได้ผลลัพธ์ที่ไม่ถูกต้อง ในหน่วยงานและบริษัทขนาดใหญ่ นั่นโอกาสที่ข้อมูลทั้งหมดจะ “สะอาด” นั้นเป็นเรื่องที่ยาก ดังนั้นจึงจำเป็นต้องมีผู้ทำหน้าที่กลั่นกรองและควบคุมคุณภาพของข้อมูลด้วย

เราสามารถสรุปความแตกต่างของคลังข้อมูลกับฐานข้อมูลที่ใช้ประจำวันได้แต่ละหัวข้อดังนี้

1. Consistency ทั้ง OLTP และ คลังข้อมูล ต่างก็ให้ความสำคัญในเรื่องข้อมูลควรจะมีคุณสมบัติเหมือนกัน สำหรับ OLTP ซึ่งมีการทำ transaction จำนวนมากซึ่งที่ต้องการคือการทำให้ครบ ไม่มีการสูญหาย ดังนั้นจึงมีความจำเป็นผู้ส่งและผู้รับจะต้องรับรู้และตรวจสอบอยู่ตลอดเวลาว่าขณะนี้มีการทำ transaction เกิดขึ้นหรือไม่ สำหรับคลังข้อมูล จะไม่สนใจทำการทำ transaction แต่จะสนใจว่าการ load data ใหม่เข้ามานั้นทำสำเร็จหรือยัง และการ load data เข้ามาทั้งหมดนั้นถูกต้องหรือไม่
2. Transaction สำหรับระบบ OLTP นั้น ในแต่ละวันอาจมีการทำ transaction มากมายซึ่งการทำ transaction แต่ละครั้งจะใช้ข้อมูลเพียงแค่เล็กน้อยเท่านั้น สำหรับคลังข้อมูล แต่ละวันจะทำแค่เพียง 1 transaction ซึ่ง transaction นี้จะต้องใช้ข้อมูลเป็นจำนวนมากมาย
3. Time Dimension สำหรับ OLTP นั้นจะทำงานอย่างรวดเร็วและทำ transaction อย่างสม่ำเสมอ สถานะของข้อมูลต่างๆมีการเปลี่ยนแปลงอยู่ตลอดเวลา และความสัมพันธ์ระหว่างเอนิตีต่างๆก็เปลี่ยนแปลงไปด้วย สำหรับระบบคลังข้อมูลมักจะเก็บข้อมูลในอดีตเพื่อใช้ในการวิเคราะห์ ดังนั้นข้อมูลจะไม่ค่อยมีการเปลี่ยนแปลงตลอดวัน

เป็นที่น่าสังเกตว่าคลังข้อมูลไม่ต้องทำการ normalization เหมือนกับฐานข้อมูลประจำวันที่ต้องทำการ normalization ทั้งนี้เพราะในฐานข้อมูลประจำวัน ข้อมูลจำนวนมากมีการเปลี่ยนแปลงทำให้ทันสมัยอยู่ตลอดเวลา ดังนั้นประเด็นสำคัญจึงอยู่ที่การเปลี่ยนแปลงทำให้ทันสมัย การออกแบบฐานข้อมูลประจำวันจึงต้องทำให้มีความซ้ำซ้อนหรือ redundancy น้อยที่สุด วิธีการที่จะทำได้ตามจุดประสงค์คือการทำ normalization แต่สำหรับข้อมูลในคลังข้อมูลเป็นข้อมูลที่มีการกลั่นกรองมาแล้ว ใช้ในการวิเคราะห์ตอบคำถามของผู้บริหาร ประเด็นสำคัญจึงไม่อยู่ที่การทำให้ทันสมัย ทำให้ข้อมูลในคลังข้อมูลสามารถมีความซ้ำซ้อนได้ เพราะความซ้ำ

ข้อนี้มีข้อดีคือ การตอบคำถามและการออกรายงานสามารถทำได้รวดเร็ว เนื่องจากไม่ต้อง join หลายตาราง ดังนั้นในคลังข้อมูลจึงไม่มีความจำเป็นต้องทำการ normalization

เรื่องที่ 15.2.2 โครงสร้างสถาปัตยกรรมและองค์ประกอบของคลังข้อมูล

จากความหมายของคลังข้อมูลที่เป็นที่เก็บข้อมูลขององค์กรที่ได้รับการออกแบบเพื่อช่วยในการตัดสินใจของฝ่ายบริหารนั้น ในทางปฏิบัติสิ่งที่เก็บอยู่ในโรงเก็บข้อมูลไม่ได้มีแต่เพียงข้อมูลเท่านั้น หากยังเก็บเครื่องมือสำหรับดำเนินการข้อมูล กระบวนการทำงานกับข้อมูล และทรัพยากรอื่นๆ ด้วยเช่น ภาพลักษณ์ของเอกสาร ภาพถ่าย แผนที่ เป็นต้น

ข้อมูลในคลังข้อมูลก็คือข้อมูลธุรกรรมของหน่วยงาน ดังนั้นจึงจำเป็นต้องมีเครื่องมือสำหรับแยกข้อมูลออกจากฐานข้อมูลองค์กรมาเก็บไว้ ข้อมูลเหล่านี้จำเป็นต้องมีเมตาเดตา (metadata) สำหรับใช้พรรณนาลักษณะของข้อมูล ต้นกำเนิด รูปแบบ ซึ่ดจำกัดในการใช้ และ ลักษณะอื่น ๆ ของข้อมูลที่ใช้ในการกำหนดว่าจะใช้ข้อมูลนั้นอย่างไร

คลังข้อมูลอาจจะมีข้อมูลเป็นจำนวนมากมายมหาศาล ดังนั้นจึงจำเป็นต้องมีฐานข้อมูลของตนเองในการเก็บและประมวลผลข้อมูล หน่วยงานต้องมีโปรแกรมจัดการฐานข้อมูลและโปรแกรมอื่นๆ สำหรับช่วยในการเปลี่ยนรูปแบบข้อมูล จัดกลุ่มข้อมูล รวมข้อมูล และโยกย้ายข้อมูลจากฐานข้อมูลหนึ่งไปยังฐานอื่นๆ โปรแกรมเหล่านี้ต้องทำงานได้ทั้งกับข้อมูลที่เป็นจำนวน ข้อมูลกราฟิก ข้อมูลภาพลักษณ์ และ ข้อมูลแบบมัลติมีเดีย โปรแกรมเหล่านี้จะต้องสามารถแปลงข้อมูลให้เหมาะสมที่จะนำไปวิเคราะห์ และจัดทำรายงานในรูปแบบต่างๆ ได้

โดยที่คลังข้อมูลมีบริการสำคัญหลายอย่างให้แก่ผู้ใช้ซึ่งอาจจะไม่ใช่ผู้เชี่ยวชาญในด้านคอมพิวเตอร์ ดังนั้นการจัดคลังข้อมูลจึงจำเป็นต้องจัดการฝึกอบรมให้แก่ผู้ใช้ด้วย นอกจากนั้นยังอาจจะต้องจัดระบบอธิบายการใช้เอาไว้ในระบบด้วยเพื่อให้ผู้ใช้สามารถเรียกคำอธิบายมาใช้เมื่อต้องการได้ ส่วนประกอบสำคัญอีกอย่างหนึ่งก็คือที่ปรึกษาเกี่ยวกับคลังข้อมูลเพื่อช่วยเหลือผู้ใช้ให้สามารถใช้ระบบได้อย่างมีประสิทธิภาพมากขึ้น

ตารางที่ 15.2.2 ส่วนประกอบของคลังข้อมูล

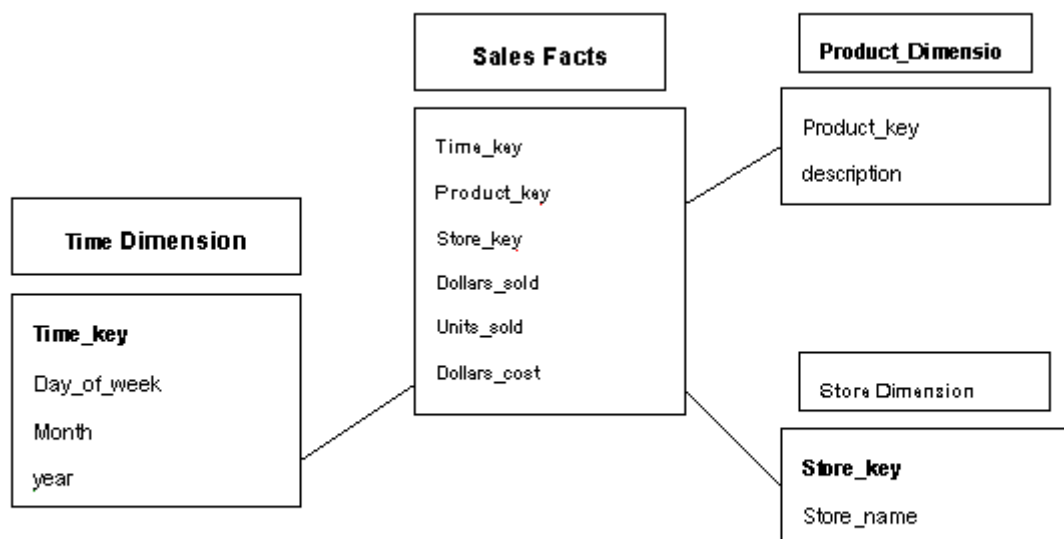
ส่วนประกอบของคลังข้อมูล
● เครื่องมือสกัดแยกข้อมูล
● ข้อมูลที่สกัดและแยกออกมาแล้ว
● เมตาเดตาสำหรับบรรยายเนื้อหาข้อมูล
● ฐานข้อมูลสำหรับคลังข้อมูล
● เครื่องมือจัดการข้อมูลในคลังข้อมูล
● โปรแกรมสำหรับจัดส่งข้อมูล
● เครื่องมือวิเคราะห์สำหรับผู้ใช้

- วัสดุและหลักสูตรการฝึกอบรม
- ที่ปรึกษาด้านคลังข้อมูล

ในที่นี้จะขอล่าวรายละเอียดของฐานข้อมูลสำหรับคลังข้อมูลซึ่งประกอบด้วย

1. **Dimensional Modeling** เป็นชื่อเรียกของเทคนิคในการทำให้ฐานข้อมูลง่ายต่อการทำความเข้าใจ โดยการมองภาพของฐานข้อมูลเป็นลูกบาศก์ที่มี 3,4,5 มิติ หรือมากกว่านั้น ทำให้สามารถจินตนาการการ หั่นหรือ แบ่งลูกบาศก์ที่มีลักษณะเหมือนลูกเต๋านี้ได้ นั่นคือสามารถตัดข้อมูลมาวิเคราะห์ดูในช่วงใดก็ได้ และหมุนข้อมูล ดูได้จากทุกๆด้านของลูกเต๋า ตัวอย่างเช่น เราขายสินค้า (product) ในหลายๆที่ (market) และในช่วงเวลาต่างๆ กัน (time) เราสามารถสร้าง Dimensional Modeling ได้โดยให้ label คือ product, market และ time อยู่บนแต่ละด้านของลูกบาศก์ที่เป็น 3 มิติ แต่ละจุดภายในลูกบาศก์เกิดจากการตัดของ coordinate ซึ่งมี label อยู่ที่ยอดของลูกบาศก์ ดังนั้นจุดต่างๆภายในลูกบาศก์คือผลลัพธ์ทางด้านธุรกิจที่พิจารณาจากทั้ง 3 เรื่องคือ สินค้า, ที่ขาย สินค้า, เวลา พร้อมๆกัน

2. **Star Join Schema** เป็นชื่อหนึ่งของ dimensional model ซึ่งเป็นชื่อที่ใช้กันมานานเนื่องจาก diagram มีรูปร่างคล้ายดาว ซึ่งมีตารางใหญ่ 1 ตารางอยู่ตรงกลางซึ่งเรียกว่า *fact table* และมีตารางเล็กๆที่มีความสัมพันธ์กับตารางหลักนั้นอยู่รอบๆ เรียกว่า *dimensional table* ซึ่งตารางหลักนี้เป็นตารางเดียวที่ใช้ multiple join เพื่อเชื่อมต่อกับตารางอื่นๆ แต่ตารางอื่นๆที่อยู่รอบๆจะมีเพียงแค่ single join เพื่อเชื่อมเข้ากับตา



รูปที่ 15.2.2 แสดงภาพ Star Join Schema

วางหลักเท่านั้น

2.1 fact table ข้อมูลที่เก็บอยู่ใน fact table เรียกว่า *grain of fact table* แต่ละเรคคอร์ดใน fact table จะแสดงถึง ยอดขายทั้งหมดของสินค้าหนึ่งที่ยาได้ในสถานที่ที่แห่งหนึ่งในหนึ่งวัน fact table จะเป็นที่ยึดข้อ

มูลที่วัดได้ของธุรกิจหนึ่งๆ เช่นข้อมูลที่เป็นตัวเลข ข้อมูลที่วัดได้แต่ละตัวนี้มาจากการ intersection กันของทุกๆ dimensions จากรูป 15.2.2 ตัวเลขที่วัดได้ที่อยู่ในตาราง sales fact คือ จำนวนเงิน (number of dollars), จำนวนหน่วยสินค้าที่ขายได้ (number of units sold), ราคาสินค้า (cost)

2.2dimensional table เก็บคำอธิบายของแต่ละ dimension ของธุรกิจเอาไว้ ซึ่งคำอธิบายเหล่านี้จะช่วยให้การอธิบายถึงสมาชิกในทุกๆ dimension และใน dimension table จะประกอบด้วยหลายๆ attributes ซึ่ง attribute ที่จะต้องเป็นตัวอักษรและแต่ละ attribute ต้องแยกออกจากกัน

เรื่องที่ 15.2.3 กระบวนการการรวบรวมข้อมูลและวิเคราะห์หารูปแบบข้อมูลในคลังข้อมูล

เนื่องจากข้อมูลพื้นฐานของฐานข้อมูลในคลังข้อมูลประกอบด้วยเรคคอร์ดจำนวนมากจะต้องมีการออกแบบคลังข้อมูลเพื่อรวบรวมข้อมูลและวิเคราะห์หารูปแบบข้อมูลในคลังข้อมูล ขั้นตอนการออกแบบประกอบด้วย 4 ขั้นตอน

ขั้นตอน 1 เลือก business process ที่ต้องการสร้างซึ่งเป็นกระบวนการหลักที่ต้องการทำในองค์กร ซึ่งกระบวนการนั้นมีระบบเดิมสนับสนุนอยู่ ข้อมูลในระบบนั้นสามารถนำมารวบรวมเพื่อทำเป็นคลังข้อมูลได้ เช่น ใบส่งของ (order), ใบส่งของ (invoices), รายการสินค้า (inventory), ยอดขาย (sales) business process ที่เลือกมานั้นต้องเป็นหัวข้อธุรกิจที่สนใจ เพื่อจะได้สามารถทำการออกแบบคลังข้อมูลให้เกี่ยวข้องกับเฉพาะหัวข้อธุรกิจที่สนใจเท่านั้น ส่วนข้อมูลที่ไม่ได้เกี่ยวข้องกับหัวข้อธุรกิจก็ไม่นำมารวมในคลังข้อมูล

ขั้นตอน 2 เลือก grain ของ business process ข้อมูลที่เป็นข้อมูลพื้นฐานที่เก็บอยู่ใน fact table เรียกว่า grain ใน business process นี้มี grain ที่มีอยู่ทั่วไป เช่น ข้อมูลของการทำ transaction ในแต่ละครั้ง (individual transaction), ข้อมูลของการทำงานในแต่ละวัน, สรุปในแต่ละวัน (individual daily snapshots), ข้อมูลจากการสรุปการทำงานในแต่ละเดือน (individual monthly snapshots)

ขั้นตอน 3 เลือก dimension ที่จะถูกนำมาใช้กับแต่ละ record ของ fact table นั้น dimension ที่มีอยู่ทั่วไปเช่น เวลา (time), สินค้า (product), ลูกค้า (customer) ซึ่งแต่ละ dimension จะถูกอธิบายแยกกัน ในลักษณะของ dimensional attribute ซึ่งอธิบายแต่ละ dimension เป็นตัวหนังสือ ซึ่ง attributes เหล่านี้จะถูกเก็บอยู่ในแต่ละตาราง dimension

ขั้นตอน 4 เลือก measured fact (ข้อมูลที่มีการวัด, การประมวลผล หรือการคำนวณไว้แล้ว) ที่จะเก็บอยู่ในแต่ละเรคคอร์ดของ fact table ปริมาณต่างๆหรือ measured fact ที่เพิ่มเข้าไปซึ่งจะมีลักษณะเป็นตัวเลข ได้แก่ ปริมาณที่ขายได้ (quantity sold) และ จำนวนเงินที่ได้รับจากการขาย (dollars sold)

ถ้านักวิเคราะห์ต้องการสอบถามข้อมูล (query) โดยไม่มีการกำหนดเงื่อนไขให้กับบางตาราง dimension แล้วในคลังข้อมูลมีเฉพาะข้อมูลพื้นฐานเท่านั้น นั้นหมายความว่า การสอบถามข้อมูล (query) จะต้องทำการรวมข้อมูลภายในเรคคอร์ดจำนวนมากมาย ถ้าเป็นเช่นนี้จะทำให้การทำ query แต่ละครั้งเกิดการสิ้นเปลืองมาก ดังนั้นจึงต้องทำการรวบรวมหรือข้อสรุป (aggregate) ข้อมูลไว้ล่วงหน้าเพื่อเร่งให้การทำ query สามารถทำได้เร็วขึ้นมีประสิทธิภาพในการทำงานมากขึ้น การรวบรวมหรือข้อสรุป (aggregate) สามารถถูกสร้างได้มากมายตามที่ต้องการในเฉพาะส่วนที่ต้องการเท่านั้น

ในการสอบถามข้อมูล (query) ในคลังข้อมูลจะใช้ Structured Query Language (SQL) เป็นมาตรฐานสำหรับการสอบถามข้อมูลทั้งหมดในคลังข้อมูลที่เกี่ยวข้องกับ fact table การสอบถามข้อมูลจะต้องมีการใช้ dimension table ในการทำการสอบถามข้อมูลเสมอ เพราะใน dimension table จะเก็บชื่อและคำอธิบายที่การสอบถามข้อมูลต้องการใช้เอาไว้ การสอบถามข้อมูลจะประกอบไปด้วย *Where Clause* 2 ส่วนคือ

- การ join ระหว่าง fact table และ dimension table
- เซตของข้อกำหนดหรือเงื่อนไข (criteria) สำหรับคอลัมน์ที่อยู่ใน dimension table

กิจกรรม 15.2

1. จงอธิบายลักษณะของการประมวลผลธุรกรรมออนไลน์ และการประมวลผลเชิงวิเคราะห์ออนไลน์
2. จงบอกส่วนประกอบของโรงเก็บข้อมูลมา 5 อย่าง

แนวตอบกิจกรรม 15.2

1. การประมวลผลธุรกรรมออนไลน์ คือ กระบวนการบันทึกและประมวลผลข้อมูลธุรกรรมในลักษณะอัตโนมัติ โดยใช้อุปกรณ์ช่วยในการบันทึกข้อมูลที่สามารถจัดเก็บข้อมูลลงในฐานข้อมูลของระบบคอมพิวเตอร์ได้ทันที

การประมวลผลเชิงวิเคราะห์ออนไลน์ คือ กระบวนการค้นคืนข้อมูลที่มีอยู่แล้วในฐานข้อมูล แล้วนำมาวิเคราะห์อย่างละเอียด เพื่อใช้เป็นข้อมูลประกอบการตัดสินใจของผู้บริหาร นักวิจัยตลาด นักสถิติ หรือผู้ใช้อื่นๆ

2. โรงเก็บข้อมูล มีส่วนประกอบที่สำคัญดังนี้
 - 2.1 เครื่องมือสกัดแยกข้อมูล
 - 2.2 เมตาเดตาสำหรับบรรยายเนื้อหาข้อมูล
 - 2.3 ฐานข้อมูลสำหรับโรงเก็บข้อมูล
 - 2.4 เครื่องมือจัดการข้อมูลในโรงเก็บข้อมูล
 - 2.5 เครื่องมือวิเคราะห์สำหรับผู้ใช้