บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1.นิยามของ คลังข้อมูล

คลังข้อมูล เป็นการแยกฐานข้อมูลที่จะใช้งานกับระบบช่วยเหลือการตัดสินใจ ออกจากระบบฐานข้อมูลประจำวัน โดยฐานข้อมูลของคลังข้อมูลอาจจะนำข้อมูลมาจากฐานข้อมูลประจำวันหรือมีการประมวลผลข้อมูลจากฐานข้อมูลประจำวันให้กลายเป็นข้อมูลสารสนเทศหรือนำข้อมูลอื่นๆ เพิ่มเติมเข้ามาก็ได้

นิยามของคลังข้อมูล คือ ฐานข้อมูลที่ช่วยในการจัดหาข้อมูล สำหรับกระบวนการในการตัดสินใจ โดยฐานข้อมูลดังกล่าวมีลักษณะดังนี้

2.1.1 Subject Oriented

คลังข้อมูลมีการจัดเก็บข้อมูลตามหัวข้อที่เราสนใจ ตัวอย่างเช่น เพื่อเรียนรู้เกี่ยวกับข้อมูลการขายของบริษัทเราก็สามารถสร้างคลังข้อมูล โดยเน้นไปที่การขายและด้วยการใช้คลังข้อมูลเราสามารถตอบคำถามได้ว่า “ใครคือลูกค้าที่ดีที่สุดของเราในปีที่แล้วได้” นี่คือความสามารถที่จะกำหนดการเก็บข้อมูลในคลังข้อมูล โดยใช้เนื้อหาหลักเป็นตัวกำหนดซึ่งในกรณีนี้คือการขาย ทำเป็นคลังข้อมูลแบบ Subject Oriented

2.1.2 Integrated

คลังข้อมูลต้องนำข้อมูลจากหลายๆแหล่งข้อมูลใส่ลงไปในโครงสร้างที่สร้างขึ้นดังนั้นข้อมูลจึงต้องถูกแก้ปัญหาในเรื่องของความขัดแย้งกันและความไม่ตรงกันระหว่างหน่วยของการวัดเมื่อทำพวกนี้สำเร็จเราจะเรียกว่า ถูก Integrated ซึ่งจะได้ข้อมูลที่สอดคล้องเป็นหนึ่งเดียว

2.1.3 Nonvolatile

ข้อมูลในคลังข้อมูล จะไม่เปลี่ยนแปลงนี่เป็นลักษณะของเชิงแนวคิด(Logical) เพราะจุดประสงค์ในการใช้ คลังข้อมูล คือเพื่ออ่านเท่านั้น

2.1.4 Time Variant

เพื่อค้นพบแนวโน้มทางธุรกิจ ผู้วิเคราะห์จำเป็นต้องมีข้อมูลจำนวนมาก และต้องมีการเก็บข้อมูลย้อนหลังเป็นเวลานาน 5 ปี ถึง 10 ปี นั่นก็หมายความว่าข้อมูลแต่ละตัวมีองค์ประกอบที่สำคัญคือเวลาที่บอกคุณลักษณะของข้อมูลนั้นอยู่

2.2 จุดมุ่งหมายของ คลังข้อมูล

เป้าหมายของการออกแบบ คลังข้อมูล คือ การแยกกลุ่มข้อมูลสารสนเทศที่ใช้ในการวิเคราะห์ทางธุรกิจออกจากฐานข้อมูลประจำวัน มาเก็บในระบบจัดการฐานข้อมูลที่มีประสิทธิภาพสูงและทำให้การเรียกใช้ข้อมูลชุดนี้ทำได้อย่างยืดหยุ่นจากเครื่องมือ (Tools) ที่อยู่บนเครื่องคอมพิวเตอร์ (desktop) ทั่วไป เพิ่มกลไกในการช่วยตัดสินใจ ปรับปรุงเวลาที่ตอบสนองอย่างรวดเร็วขึ้นอย่างมาก และผู้ใช้งานสามารถเรียกดูข้อมูลและรายละเอียดที่จำเป็นที่ถูกเก็บมาก่อนหน้านี้เพื่อช่วยให้การตัดสินใจทางธุรกิจแม่นยำขึ้น

2.2.1 เป้าหมายในการจัดทำคลังข้อมูล

2.2.1.1 คลังข้อมูลทำให้สามารถเข้าถึงข้อมูลขององค์กรได้ ผู้จัดการและนักวิเคราะห์ขององค์กรสามารถเชื่อมต่อเข้าไปยังคลังข้อมูล จากเครื่องคอมพิวเตอร์ของตนได้ ซึ่งการเชื่อมต่อนี้สามารถทำได้ในทันทีตามความต้องการและมีประสิทธิภาพสูงและเครื่องมือ (Tools) ที่มีให้กับผู้จัดการและนักวิเคราะห์สามารถใช้งานได้สะดวกและง่าย

2.2.1.2 ข้อมูลในคลังข้อมูล จะต้องถูกต้องตรงกันหมด คำถามเดียวกันต้องได้รับคำตอบที่เหมือนกันเสมอ ไม่ว่าผู้ถามจะเป็นใคร ถามเวลาใดก็ตาม

2.2.1.3 ข้อมูลในคลังข้อมูล สามารถถูกตัดและหมุนดูได้ทุกแกน หมายถึง ข้อมูลสามารถถูกวิเคราะห์จากหัวข้อในธุรกิจประเภทนั้น โดยแบ่งข้อมูลหรือรวมข้อมูลมาวิเคราะห์ตามต้องการ

2.2.1.4 คลังข้อมูล เป็นส่วนที่ผลิตข้อมูลจากระบบ OLTP (Online Transaction Processing) ข้อมูลไม่เพียงแต่ถูกรวบรวมมา ไว้ที่ศูนย์กลางอย่างเดียว แต่จะถูกรวบรวมอย่างระมัดระวังจากแหล่งข้อมูลหลายๆแหล่งภายนอกองค์กรด้วย แล้วนำมาปรับปรุงให้เหมาะสมกับการใช้งานเท่านั้น ถ้าข้อมูลเชื่อถือไม่ได้หรือไม่สมบูรณ์ก็จะไม่ได้รับอนุญาตให้นำไปใช้ได้

2.3 ระบบฐานข้อมูลแบบทรานแซคชั่น (Transaction)

ตามธรรมชาติของระบบฐานข้อมูลประจำวันจะเป็นระบบทรานแซคชั่นแต่เนื่องจากทรานแซคชั่นมักจะต้องเรียกใช้ข้อมูลจากฐานข้อมูลขนาดใหญ่ เมื่อต้องการเรียกใช้ข้อมูลเพื่อช่วยในการตัดสินใจ ทำให้มีข้อมูลจำนวนมหาศาล จึงมีปัญหาในการเรียกข้อมูลทรานแซคชั่น ดังนี้

2.3.1 ข้อมูลทรานแซคชั่นเรียกใช้ได้ยาก

2.3.2 มีการแตกตารางที่นอร์มอลไลซ์ (Normalize)แล้ว ออกเป็นหลายตาราง

2.3.3 ไม่รองรับคำถามที่สนับสนุนการตัดสินใจเพราะคำถามทั้งหมดของระบบสนับสนุนการตัดสินใจมีการรวมกันของตารางต่างๆที่ซับซ้อน ซึ่งทำให้ประสิทธิภาพของการทำงานของระบบทรานแซคชั่นลดลงและทำงานช้าลงไป

2.3.4 มีการเก็บข้อมูลย้อนหลังน้อย

ดังนั้นเมื่อ คลังข้อมูล ถูกคิดขึ้นมา เพื่อช่วยให้ผู้ใช้เรียกใช้ข้อมูลด้วยวิธีที่สร้างสรรค์ เพราะธรรมชาติที่แตกต่างกันของคลังข้อมูล และฐานข้อมูลประจำวัน โดยคลังข้อมูล จะแยกฐานข้อมูลออกมา เก็บข้อมูลสารสนเทศที่สรุปมาเฉพาะหัวข้อที่สนใจ เพื่อวิเคราะห์ใช้ในการบริหารและควบคุมธุรกิจ โดยระบบนี้พิจารณาถึงการใช้ประโยชน์จากหัวข้อเป็นหลัก

2.4 ความแตกต่างของ คลังข้อมูล กับระบบฐานข้อมูลประจำวัน

คลังข้อมูล และระบบฐานข้อมูลประจำวันนั้นมีความแตกต่างกันอย่างมากมาย ดังนี้

2.4.1 ความถูกต้องตรงกัน (Consistency)

ทั้งระบบ OLTP (Online Transaction Processing) และคลังข้อมูล ต่างก็ให้ความสำคัญในเรื่องของความถูกต้องตรงกันของข้อมูล (Data consistency) สำหรับ OLTP (Online Transaction Processing) ซึ่งมีการทำทรานแซคชั่นจำนวนมากๆนั้น สิ่งที่ต้องการคือ การทำทรานแซคชั่นให้ครบ (ไม่มีการสูญหาย) ดังนั้นจึงมีความจำเป็นที่ผู้ส่งและผู้รับจะต้องรับรู้และตรวจสอบอยู่ตลอดเวลาว่าขณะนี้มีการทำทรานแซคชั่นเกิดขึ้นหรือไม่

สำหรับคลังข้อมูล จะไม่สนใจการทำทรานแซคชั่นแต่ละครั้ง แต่จะสนใจว่าการนำข้อมูลใหม่เข้ามานั้นทำเสร็จแล้วหรือยัง (นำข้อมูลเข้ามาทั้งหมดและข้อมูลถูกต้อง)

2.4.2 ทรานแซคชั่น (Transaction)

ระบบ OLTP (Online Transaction Processing) นั้นในแต่ละวันอาจมีการทำทรานแซคชั่นเป็นหมื่นเป็นแสนครั้ง ซึ่งการทำทรานแซคชั่นแต่ละครั้งจะใช้ข้อมูลเพียงเล็กน้อยเท่านั้น

สำหรับคลังข้อมูล แต่ละวันจะทำเพียงแค่ 1 ทรานแซคชั่น ซึ่งการทำทรานแซคชั่นแต่ละครั้งอาจต้องใช้ข้อมูลเป็นล้านเรคอร์ด (Record) เลยก็ได้ ดังนั้นเราจึงเรียกกระบวนการนี้ว่า Production Data Load แทนและสิ่งที่เราสนใจในกระบวนการนี้มีเพียงแค่ Production Data Load เท่านั้น ถ้าการทำ Production Data Load ถูกทำให้หยุดกลางคัน ระบบจะทำการเอาข้อมูลที่เคยมีมาก่อนที่ Production Data load จะทำ มาเขียนทับลงทันที

2.4.3 ผู้ใช้และผู้จัดการ

สำหรับ OLTP (Online Transaction Processing) นั้น ผู้ใช้ คือ ผู้ที่ทำงานอยู่ในองค์กรนั้น (อาจเป็นเจ้าหน้าที่ที่ทำหน้าที่รับออร์เดอร์ (order), เจ้าหน้าที่ที่ทำหน้าที่ตรวจสอบและดูแลเงิน, เจ้าหน้าที่ที่ทำหน้าที่คอยบริการลูกค้าใหม่, เจ้าหน้าที่ที่ทำหน้าที่ป้อนข้อมูลเข้าไป) ซึ่งส่วนใหญ่ผู้ใช้เหล่านี้ ในช่วงเวลาหนึ่งจะทำงานกับ 1 รายการเท่านั้น ผู้ใช้ OLTP (Online Transaction

Processing) มักจะทำงานที่ลักษณะเป็นงานซ้ำๆ เดิมๆ รายงานส่วนใหญ่ที่ได้จากการทำงานบนระบบ OLTP (Online Transaction Processing) นี้มักจะมีลักษณะเป็นลิสต์ (List) ของตารางเลย

สำหรับคลังข้อมูล ผู้ใช้คือผู้ที่ทำหน้าที่คอยดูแลการทำงานของพนักงานในองค์กรเท่านั้น (คอยนับจำนวนออร์เดอร์ใหม่ๆ, หาเหตุผลว่าทำไมลูกค้าจึงไม่พอใจ, คอยตรวจสอบดูว่าข้อมูลอะไรใหม่ๆ เข้ามาบ้าง, คอยตรวจสอบและแก้ไขข้อผิดพลาดของข้อมูล) ผู้ใช้ของคลังข้อมูล จะไม่ทำงานทีละ 1 รายการแต่จะพิจารณาจากรายการทั้งหมด แล้วหาคำตอบที่ต้องการออกมา (Answer Set ขนาดเล็กๆ) และคำถามที่ให้กับคลังข้อมูลก็สามารถเปลี่ยนแปลงได้เรื่อยๆ (ไม่ใช่คำถามเดิมที่ถามก็ได้)

2.4.4 Time ไดเมนชัน

OLTP (Online Transaction Processing) จะทำงานอย่างรวดเร็วและทำทรานแซคชั่นอย่างสม่ำเสมอ (การวัดเวลาใช้หน่วยเป็นนาทีและวินาที) สถานะของข้อมูลต่างๆ มีการเปลี่ยนแปลงอยู่ตลอดเวลา และความสัมพันธ์ระหว่างเอนติตี้ (Entity) ต่างๆ ก็จะเปลี่ยนแปลงไปด้วย

ฐานข้อมูลใน OLTP (Online Transaction Processing) จะขาดการสนับสนุน (การอ้างอิง) จากข้อมูลในอดีต เพราะในคลังข้อมูล มักจะมีคำถามถึงการวิเคราะห์ข้อมูลในอดีต แม้ว่าเราสามารถเก็บข้อมูลในอดีตไว้ใน OLTP (Online Transaction Processing) ได้ แต่ก็ต้องเป็นภาระหนักของระบบในการทำให้มองเห็นภาพในอดีต และจะทำให้ยากในการทำทรานแซคชั่น (ข้อมูลมีมากขึ้นก็ต้องใช้เวลามากขึ้นด้วย)

คลังข้อมูล จะต้องไม่ถูกเปลี่ยนแปลงตลอดวัน ดังนั้นการนำข้อมูลเข้าออกตลอดทั้งวันก็จะไม่มีทางเกิดขึ้นและมีการระมัดระวังในการเก็บข้อมูลลงไปใน คลังข้อมูลแต่ละครั้งด้วย

2.4.5 การสร้างข้อมูล (Data Model)

ความแตกต่างที่สำคัญที่สุดของ OLTP (Online Transaction Processing) และ คลังข้อมูล คือ การสร้างและออกแบบคลังข้อมูลหรือ Data Model นั่นเอง

OLTP (Online Transaction Processing) จะใช้ Entity Relation Modeling คือการกำจัดความซ้ำซ้อนให้หมดไปเพื่อการทำทรานแซคชั่นจะได้สามารถทำได้ง่ายและรวดเร็วยิ่งขึ้น (สามารถทำทรานแซคชั่นได้โดยเข้าถึงข้อมูลเพียงตัวเดียวเท่านั้น) การทำ Entity Relation Modeling นั้น จะใช้วิธีแยกข้อมูลออกเป็นตารางเล็กๆ และแต่ละตารางสามารถเชื่อมต่อไปยังตารางอื่นได้ ซึ่งข้อสังเกตเกี่ยวกับ OLTP (Online Transaction Processing) จะใช้ Entity Relation Diagram ซึ่งมีลักษณะดังนี้

2.4.5.1 ไดอะแกรมนี้ซิมเมทริก (symmetric) หมายถึง ทุกๆ ตารางเหมือนกัน

หมด ไดอะแกรมนี้ไม่สามารถบอกได้ว่าตารางไหนสำคัญกว่าหรือใหญ่กว่า และไม่สามารถบอกได้ว่า ตารางใดบรรจุตัววัดที่เป็นชนิดตัวเลข (Numerical Measurement) ทางธุรกิจ ซึ่งไดอะแกรมรูปแบบนี้ ผู้ใช้จะทำความเข้าใจและจดจำได้ยาก

2.4.5.2 ถ้ามี 2 ตารางในไดอะแกรมต้องการที่จะรวมกันก็มีหลายวิธีที่จะทำได้ และไม่ว่าวิธีไหนก็จะได้คำตอบเดียวกัน แต่ในระหว่างการทำ Inner join นั้นจะมีการใช้องค์ประกอบของข้อมูล (Data Element) ไม่เหมือนกัน

สำหรับคลังข้อมูล จะใช้ Dimensional Model หรือ Star Join Schema ซึ่งเป็นชื่อ ที่นักออกแบบฐานข้อมูลใช้กันมานานแล้ว เนื่องจากไดอะแกรมมีรูปร่างคล้ายดาว ซึ่งมีตารางหลัก 1 ตาราง อยู่ตรงกลางและมีตารางเล็กๆ ที่มีความสัมพันธ์กับตารางหลักนั้นอยู่รอบๆ ซึ่งตารางหลักนี้เป็นตารางเดียวที่ใช้ร่วมกับตารางอื่นแบบ Multiple join เพื่อเชื่อมเข้ากับตารางอื่นๆ แต่ตารางที่อยู่รอบๆนั้นจะมีเพียง Single join เพื่อเชื่อมเข้ากับตารางหลักเท่านั้น โดยตารางหลัก เรียกว่า ตารางแฟ็ค (Fact table) ส่วนตารางอื่นๆ จะเรียกว่าตารางไดเมนชัน (Dimension table)

2.4.6 The Standard Template Query

คลังข้อมูล จะใช้ SQL เป็น template มาตรฐานสำหรับการตั้งคำถามทั้งหมด ใน คลังข้อมูล ซึ่งจะเกี่ยวกับ Fact table ซึ่ง template ของ SQL ประกอบด้วย

2.4.6.1 Select list : จะเลือกคอลัมน์ที่ต้องการให้ปรากฏในเซตคำตอบของผู้ใช้ไว้ ซึ่งปกติแล้วใน Select list เราจะจัดชื่อหัวข้อรายการไว้เป็นรายการแรกใน Select list

2.4.6.2 From clause : จะได้มาจากชื่อของตารางที่จะต้องใช้ในคำถาม

2.4.6.3 Where clause : การทำงานเงื่อนไขของการรวม (join constrain) ซึ่งการใช้คำสั่งนั้นจะเขียนถึงความสัมพันธ์ระหว่างตารางแฟ็คและตารางไดเมนชัน โดยคีย์หลัก (Primary key) ของตารางไดเมนชันจะมีสถานะเป็น Foreign key ของตารางแฟ็ค (Fact table)

2.4.6.4 Group by clause : บอก SQL ให้สรุปข้อมูลตามหัวข้อรายการซึ่งหัวข้อรายการนี้ปรากฏอยู่ใน Select list นั่นเอง

2.4.6.5 Order by clause : ทำการตัดสินใจว่าจะทำการเรียงลำดับค่าในเซตคำตอบให้ผู้ใช้เห็นอย่างไร

2.5 ความแตกต่างของ คลังข้อมูล กับ Operational Data Stores

คลังข้อมูล นั้นแตกต่างจากระบบฐานข้อมูลปฏิบัติการหลายส่วน ส่วนหนึ่งที่แตกต่างกันก็คือ ข้อมูลที่จัดเก็บ ในระบบปฏิบัติการ (OLTP (Online Transaction Processing) System) จะเรียกข้อมูลว่า “ข้อมูลเชิงปฏิบัติการ” หรือ “Operational Data” ส่วนใน คลังข้อมูล จะ

เรียกข้อมูลว่า “ข้อมูลที่ช่วยในการตัดสินใจ” หรือ “Dimensional Support Data” ความแตกต่างของทั้งสองระบบนี้แสดงในตารางที่ 2.1

ตารางที่ 2.1 แสดงความแตกต่างระหว่าง Operational Data กับ Decision Support Data

Operational Data	Decision Support Data
จัดการข้อมูลระหว่าง application กับ application	จัดการข้อมูลตามหัวข้อเรื่องที่ต้องการ
เป็นข้อมูลแบบบอกรายละเอียด	เป็นข้อมูลแบบสรุป
ข้อมูลสามารถเปลี่ยนแปลงได้	ข้อมูลเปลี่ยนแปลงไม่ได้
ข้อมูลมีการเคลื่อนไหวตลอดเวลา	ข้อมูลจะคงที่จนกว่าจะมีการปรับปรุงใหม่
ประสิทธิภาพเป็นสิ่งสำคัญ เพราะต้องรองรับการเข้าใช้งานพร้อมๆกันของผู้ใช้จำนวนมาก	ประสิทธิภาพยืดหยุ่นกว่า เพราะในเวลาเดียวกันจะมีผู้ใช้ค่อนข้างน้อยเข้าใช้ข้อมูล
ใช้ข้อมูลในปัจจุบัน	ใช้ข้อมูลในช่วงเวลาที่ผ่านมาแล้ว
แหล่งข้อมูลมาจากภายในองค์กร	แหล่งข้อมูลมาจากทั้งภายในและภายนอกองค์กร
ขนาดข้อมูลประมาณกิกะไบต์	ขนาดข้อมูลอยู่ระหว่างกิกะไบต์ถึงเทราไบต์
การประมวลผลแต่ละครั้ง มักมีข้อมูลไม่มากนัก เป็นเพียงยูนิต หรือ เรคอร์ด	การประมวลผลครั้งหนึ่งๆ จะมีข้อมูลจำนวนมาก เช่น หลายๆ เรคอร์ด
ใช้เวลาประมวลผลน้อย เพียงไม่กี่วินาที	ใช้เวลาประมวลผลไม่แน่นอนตั้งแต่วินาทีถึงนาที
Subject Oriented : ข้อมูลสารสนเทศสำหรับการทำธุรกิจ จะถูกจัดวางตามหัวข้อที่องค์กรนั้นสนใจ เช่น หัวข้อของสินค้าอุปโภคและบริโภคอาจเป็นผลิตภัณฑ์ , ลูกค้า , สื่อโฆษณาและเวลา ซึ่งข้อมูลจาก คลังข้อมูล นั้นจะมาจากแหล่งข้อมูลที่ยืนยันว่าถูกต้องแล้ว Application Oriented : เช่น application ด้านการสั่งซื้อสินค้า ฐานข้อมูลก็จะรวมเอาข้อมูลมาใส่ในแบบฟอร์มที่มีอยู่ ข้อมูลในอดีต : เป็นสิ่งที่ระบบฐานข้อมูลประจำวันใช้น้อยมาก เพราะมีจุดประสงค์เพื่อจับความเป็นไปได้ในปัจจุบันมากกว่า แต่ในทางธุรกิจจะใช้ข้อมูลในอดีตนี้ เพื่อหาความเป็นไปได้และแนวโน้มในอนาคตได้ ข้อมูลคงที่ : เนื่องจากการวิเคราะห์และคำนวณทางธุรกิจ จะทำได้ต่อเมื่อข้อมูลมีปริมาณคงที่เท่านั้น แต่ข้อมูลจากฐานข้อมูลประจำวันมีการเคลื่อนไหวตลอดเวลา (Dynamic) ซึ่งเป็นเหตุผลหนึ่งที่ทำให้ข้อมูลใน คลังข้อมูล ไม่มีการ update โดยตรง แต่จะใช้ข้อมูลที่ประมวลผลใหม่ซ้อนทับขึ้นไปเรื่อยๆ	

2.6 ความแตกต่างของระบบจัดการฐานข้อมูลแบบ Multidimensional และ Relational

2.6.1 การตอบคำถามของระบบจัดการฐานข้อมูลแบบ Multidimensional จะสามารถทำได้ง่ายกว่าแบบ Relational

2.6.2 ระบบจัดการฐานข้อมูลแบบ Multidimensional ใช้ดูแนวโน้มของ Attribute ว่าค่าของ Attribute นั้นมีทิศทางการเปลี่ยนแปลงในช่วงเวลาที่เราสนใจเป็นอย่างไร เพราะฐานข้อมูลแบบ Dimensional สามารถสรุปออกมาให้อยู่ในรูปแบบที่เข้าใจได้ง่ายและรวดเร็วกว่า เช่น ในการค้นหาที่มีการแสดงผลข้อมูลเกินกว่า 1 หน้าจอ ผู้ใช้ที่ใช้ฐานข้อมูลแบบ Relational จะรู้สึกไม่สะดวกในการดูแนวโน้มของข้อมูลเท่ากับการใช้ฐานข้อมูลแบบ Multidimensional (เพราะต้องทำการเลื่อนหน้าจอ)

2.6.3 ระบบจัดการฐานข้อมูลแบบ Relational จะแสดงข้อมูลระดับเล็กที่สุด ไม่สามารถย่อยได้อีก และเป็นข้อมูล ณ เวลาใดเวลาหนึ่ง แต่ระบบจัดการฐานข้อมูลแบบ Multidimensional จะแสดงข้อมูลโดยรวม ซึ่งเน้นผลสรุปของข้อมูลและมีประโยชน์ในการวิเคราะห์ข้อมูลนั้นๆ

2.6.4 ฐานข้อมูลแบบ Relational จะทำการ Normalization ซึ่งการ Normalization นั้นเป็นขั้นตอนการออกแบบฐานข้อมูล โดยใช้เทคนิค Decomposition ซึ่งจะแตก Attribute ของโครงสร้างที่ไม่พึงประสงค์ไปเป็น Relation Schema ที่เล็กกว่า ซึ่งมีคุณสมบัติตามที่ต้องการตาม Normal Form ขั้นต่างๆ ดังนี้

1NF : A relation *R* is in *first normal form* (1NF) if and only if all underlying domains atomic values only หมายความว่า ความสัมพันธ์ R จะอยู่ใน normal form ขั้นที่ 1 ได้ก็ต่อเมื่อ attribute ทุกตัวในโดเมนมีค่าเพียงค่าเดียวเท่านั้น

2NF : A relation *R* is in *second normal form* (2NF) if and only if it is in 1NF and every nonkey attribute is fully dependent on the primary key หมายความว่า ความสัมพันธ์ R จะอยู่ใน normal form ขั้นที่ 2 ได้ก็ต่อเมื่อ ความสัมพันธ์ R อยู่ใน normal form ขั้นที่ 1 แล้วและทุกๆ nonkey attributes จะไม่ขึ้นอยู่กับคีย์หลัก (ไม่ขึ้นกับสับเซตคีย์หลัก)

3NF : A relation *R* is in *third normal form* (3NF) if and only if it is in 2NF and every nonkey attribute is nontransitively dependent on the primary key หมายความว่า ความสัมพันธ์ R จะอยู่ใน normal form ขั้นที่ 3 ได้ ก็ต่อเมื่อ ความสัมพันธ์ R อยู่ใน normal form ขั้นที่ 2 แล้วและทุกๆ nonkey attribute ไม่ขึ้นต่อกันเอง

Boyce / Codd Normal Form (BCNF) : A Relation R is in *Boyce-Codd Normal Form* if and only if every determinant is a candidate key หมายความว่า ความสัมพันธ์ R จะอยู่ใน BCNF ก็ต่อเมื่อ determinant ทุกๆตัวเป็น candidate key

จุดประสงค์ของการทำนอร์มอลไลซ์ คือ เพื่อสร้างแบบจำลองความสัมพันธ์ (Relation Schema) ซึ่งเก็บข้อมูลโดยปราศจากความซ้ำซ้อน (Redundancy) ที่ไม่จำเป็น เพราะฐานข้อมูลที่ดีควรจะมีความซ้ำซ้อนน้อยที่สุดเนื่องจากความซ้ำซ้อนมีผลเสีย คือ สิ้นเปลืองเนื้อที่, สิ้นเปลืองค่าใช้จ่ายในการป้อนข้อมูล (Data entry cost) และทำให้เกิดความซ้ำซ้อนในการอัพเดท ซึ่งนำไปสู่ปัญหา update anomalies ดังต่อไปนี้

- Insert Anomalies
  - insert fact ซึ่งขัดแย้งกับ fact เดิม
  - insert fact บาง fact ไม่ได้
- Deletion Anomalies
  - การลบ Fact หนึ่งออก อาจทำให้ fact อื่นถูกลบไปด้วย
- Modification Anomalies
  - เนื่องจาก Fact ซ้ำซ้อน เมื่อเปลี่ยนแปลงค่าใน Attribute หนึ่งเราจะต้องอัพเดทหลายที่ (Multiple update) ให้ครบ ซึ่งอาจทำให้ข้อมูลเกิดความขัดแย้ง (Inconsistent) กันได้

แต่สำหรับ คลังข้อมูล นั้น ข้อมูลจะเป็นข้อมูลที่ได้รับการกลั่นกรองมาแล้ว เพื่อใช้ในการวิเคราะห์และตอบคำถามผู้บริหาร ประเด็นสำคัญจึงไม่ได้อยู่ที่การอัพเดทข้อมูล จึงทำให้ข้อมูลในคลังข้อมูล สามารถมีความซ้ำซ้อนได้ เพราะความซ้ำซ้อนก็มีข้อดีสำหรับการตอบคำถามและการสร้างรายงาน คือ สามารถทำได้เร็ว เนื่องจากไม่ต้องมีการรวมกันของหลายตาราง ดังนั้นในคลังข้อมูล จึงไม่มีความจำเป็นต้องทำการนอร์มอลไลซ์ (Normalize)

2.7 Data Warehouse และ OLAP

คลังข้อมูลคือ ระบบฐานข้อมูลที่บรรจุข้อมูลที่แสดงถึงประวัติของการดำเนินธุรกิจขององค์กรนั้นๆประวัติของข้อมูลหรือข้อมูลในอดีตถูกนำมาใช้สำหรับการวิเคราะห์และสนับสนุนการตัดสินใจหลายๆระดับเพื่อวางแผนและพัฒนาประสิทธิภาพขององค์กร ข้อมูลในคลังข้อมูลถูกจัดการในการวิเคราะห์มากกว่าระบบ OLTP (Online Transaction Processing)

โดยปกติแล้วคลังข้อมูลจะถูกนำมาใช้เป็นพื้นฐานสำหรับระบบสนับสนุนการตัดสินใจ (Decision Support System) เนื่องจากคลังข้อมูลได้ถูกออกแบบมาเพื่อแก้ปัญหาที่เกิดขึ้นเมื่อเราใช้ฐานข้อมูลของระบบ OLTP (Online Transaction Processing) มาทำการวิเคราะห์

เทคโนโลยี OLAP (OnLine Analytical Processing) ทำให้เราสามารถนำข้อมูลไปใช้ได้อย่างมีประสิทธิภาพมากยิ่งขึ้นเหมาะสำหรับการวิเคราะห์แบบออนไลน์ (online analysis) ทำให้สามารถตอบคำถามที่มีความซับซ้อนได้อย่างรวดเร็ว เทคนิคมัลติไดเมนชันแนลโมเดล (Multidimensional data model) ของ OLAP และเทคนิคการรวมค่าข้อมูล (Data Aggregation) จะเข้ามาช่วยจัดการสรุปค่าของข้อมูลที่มีปริมาณมากให้มีประสิทธิภาพยิ่งขึ้น ทำให้เราสามารถ

ประเมินได้เร็วขึ้นโดยใช้การวิเคราะห์ออนไลน์และเครื่องมือที่เป็นกราฟฟิคที่สนับสนุนระบบ OLAP ทำให้เกิดความเร็วและความยืดหยุ่นที่สนับสนุนการวิเคราะห์แบบเรียลไทม์ (real time)