บทที่ 2
แนวคิด ทฤษฎีและเอกสารที่เกี่ยวข้อง

ในการค้นคว้าอิสระเรื่อง “การพัฒนาฐานข้อมูล เพื่อจัดทำ สารสนเทศสนับสนุนการตัดสินใจของผู้บริหารด้านเทคนิค พื้นที่ภาคเหนือตอนบน บริษัท ทีทีแอนด์ที จำกัด (มหาชน) ” ผู้ค้นคว้าได้ศึกษาจากเอกสารและงานวิจัยที่เกี่ยวข้อง ซึ่งสรุปสาระสำคัญได้ดังนี้

2.1 การคัดแยกข้อมูล

Raymond J. Mooney (2009) ได้อธิบายในเอกสารการสอนไว้ว่า Information Extraction (IE) หมายถึงกระบวนการในการสกัดสารสนเทศ ออกจากเอกสารที่เราสนใจ ตัวอย่างเช่น หากเราอ่านหนังสือพิมพ์ 1 เล่ม เราคงเลือกอ่านเฉพาะคอลัมน์ที่สนใจ และในคอลัมน์ที่เราสนใจนั้น อาจยาวมาก ซึ่งหลายท่านอาจจะอ่านเฉพาะย่อหน้า หรือจุดที่สนใจเท่านั้น ดังนั้น หากเราจะทำอย่างไรให้คอมพิวเตอร์เข้าใจและสกัดเฉพาะสารสนเทศ ที่เราสนใจ เราอาจต้องใช้เทคนิค หรือ Algorithm ที่จะทำให้คอมพิวเตอร์เข้าใจเนื้อหาข้อเอกสารข้อความ และสกัดเอาสิ่งที่เราสนใจออกมา

หลักการของ IE นั้น ก็เพื่อสกัดสารสนเทศ หรือข้อมูลที่เราสนใจ ออกจากเอกสารประเภทตัวอักษร ทั้งในเอกสารที่มีรูปแบบมีโครงสร้างบางส่วน (Semi-structure) และเอกสารที่ไม่มีโครงสร้าง (Unstructured) โดยทำการแปลงเอกสารที่ได้กล่าวมาแล้ว ไปสู่ฐานข้อมูลที่มีโครงสร้าง ตัวอย่างงานด้านต่าง ๆ ที่นำ IE ไปใช้ดังนี้

1) Newspaper articles หาสารสนเทศที่สนใจ จากหนังสือพิมพ์โดยอัตโนมัติ
2) Web pages หาสารสนเทศที่สนใจจาก WebPage ตัวอย่างเช่น Google Bot เป็นต้น
3) Newsgroup messages หาสารสนเทศจากข้อความในกลุ่ม เช่น สกัดหาข้อมูลจาก Mail เพื่อหาข้อมูลที่สนใจ
4) Classified ads หาสารสนเทศเฉพาะโฆษณาที่สนใจจากโฆษณาทั้งหมด

ตัวอย่างเช่น CIA และ NSA ใช้ IE ในการสกัดหาแนวโน้ม หรือการส่งสัญญาณ จากตัวอักษรผ่านทางหนังสือพิมพ์ทุกฉบับ เว็บไซต์ที่ต้องสงสัย หรือเอกสารอื่น ๆ เพื่อหาแนวโน้มการโจมตีจากจากผู้ก่อการร้าย เป็นต้น

การ Extraction เอกสาร HTML นั้น อาจเรียกอีกอย่างหนึ่งได้ว่า Web Extraction ซึ่งส่วนใหญ่จะเป็นการสกัดสารสนเทศจากเอกสาร HTML (Semi-structured) หรืออาจเรียกได้ว่า wrapper หรือ screen scraping

Website ที่จะถูกสกัดสารสนเทศได้นั้น จะต้องมี Template ที่ชัดเจน เช่น HTML จะใช้เทคนิค regular expression pattern ในการสกัดเอกสาร เช่น

Amazon list price:
Pre-filler pattern: “List Price: ”
Filler pattern: “$d+(.d{2})?b”

แต่ปัญหาที่มักจะเกิดขึ้นนั้นคือ ภาษาของเอกสารนั้นๆ หากเป็นภาษา English จะมีข้อดี คือ คำแต่ละคำจะถูกแบ่งด้วยช่องว่าง (Space) แต่สำหรับภาษาไทยนั้นเป็นเรื่องยาก เนื่องจากเป็นภาษาที่เขียนติดๆ กันจึงทำให้ต้องใช้หลักการ NLP (Natural Language Processing) เข้ามาช่วย ตัวอย่างเช่นคำว่า ตากลม (ตา-กลม) กับ ตากลม (ตาก-ลม) นั้น เป็นเรื่องยากที่จะให้คอมพิวเตอร์เข้าใจ จึงต้องใช้หลักการทาง NLP เข้ามาช่วย อาจใช้ n-Grams หรือการดูคำใกล้เคียงจาก Corpus หรือใช้เทคนิค Dictionary ฯลฯ เพื่อให้สามารถสกัดเป็นคำภาษาไทยได้อย่างถูกต้องและถูกความหมาย

2.2 คลังข้อมูล

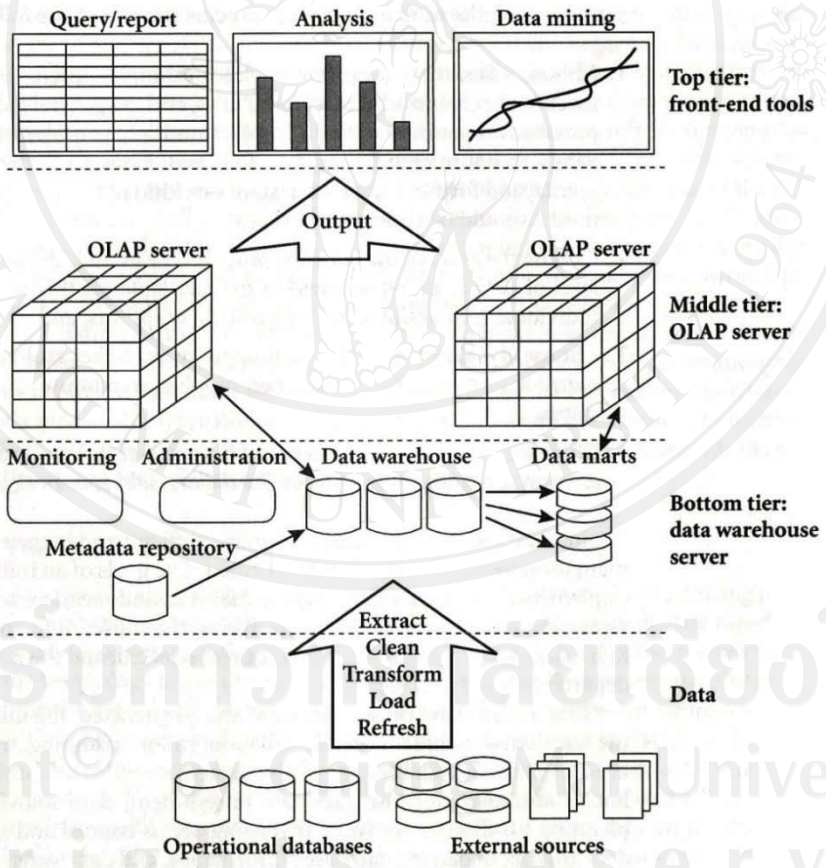
Han, Jiawei. and Kamber, Micheline. (2001) ได้อธิบายว่า คลังข้อมูล หมายถึงฐานข้อมูลเชิงสัมพันธ์ที่ถูกออกแบบให้เก็บข้อมูลการจัดการ และมีคุณสมบัติต่างๆ เพื่อใช้สำหรับช่วยสนับสนุนการตัดสินใจ โดยคลังข้อมูลมีคุณลักษณะที่ต่างจากฐานข้อมูลเชิงปฏิบัติการ เช่น หน้าที่และการจัดเก็บที่แบ่งแยกกันอย่างชัดเจน ดังนี้

1) Subject-Oriented เป็นการมองข้อมูลที่มุ่งเน้นประเด็นสำคัญๆ ของข้อมูล เช่น ข้อมูลลูกค้า ข้อมูลสินค้า ข้อมูลการขาย มากกว่าการที่จะมุ่งเน้นในงานที่มีการประมวลผลแบบวันต่อวัน คลังข้อมูลจะมุ่งเน้นในเรื่องของโครงสร้างสำหรับการวิเคราะห์ เพื่อสนับสนุนการตัดสินใจ
2) การรวบรวมข้อมูล เป็นการรวบรวมข้อมูลจากหลายแหล่งข้อมูลที่แตกต่างกัน เช่น ฐานข้อมูลเชิงสัมพันธ์ เท็กซ์ไฟล์ และออนไลน์ทรานเซ็กชันเรคอร์ด เป็นต้น โดยอาศัย

เทคนิคการรวบรวม และทำให้ข้อมูลสมบูรณ์ เพื่อให้ได้โครงสร้างข้อมูลมีความมั่นคง น่าเชื่อถือ เพื่อสามารถนำไปใช้ในการสนับสนุนการตัดสินใจได้

3) Time-Variant ข้อมูลที่จัดเก็บในคลังข้อมูลเป็นสารสนเทศที่ได้จากอดีต ซึ่งเวลาเป็นปัจจัยที่สำคัญของโครงสร้างคลังข้อมูล ทั้งในส่วนที่แสดงนัย และแสดงอย่างชัดเจน

4) Nonvolatile เป็นการพิจารณาแหล่งจัดเก็บคลังข้อมูล ที่จัดเก็บแยกจากฐานข้อมูล ทรานเซ็กชัน ซึ่งการแยกแหล่งจัดเก็บนี้มีผลทำให้คลังข้อมูลไม่จำเป็นต้องมีการประมวลผล ทรานเซ็กชันทุกครั้งที่มีการเรียกค้นหา และข้อมูลมีลักษณะอ่านได้ อย่างเดียว โดยจะมีกระบวนการที่สำคัญคือ การอ่านข้อมูลเข้าสู่หน่วยความจำ และการเข้าถึงข้อมูลในหน่วยความจำ



รูป 2.1 แสดงสถาปัตยกรรมคลังข้อมูล

ที่มา: Han, Jiawei. and Kamber, Micheline. (2001 : Figure 2.12)

สถาปัตยกรรมของคลังข้อมูล ดังแสดงในรูป 2.1 แบ่งเป็น 3 ชั้นคือ

1) ชั้นล่าง (Bottom Tier) เป็นเซิร์ฟเวอร์คลังข้อมูล ส่วนใหญ่จะเป็นระบบ ฐานข้อมูลเชิงสัมพันธ์ ข้อมูลของคลังข้อมูลได้มาจากฐานข้อมูลปฏิบัติการ (Operational Database) และแหล่งข้อมูลภายนอกผ่านโปรแกรมเกตเวย์ (Gateways) สนับสนุนการทำงานของระบบการจัดการฐานข้อมูล และอนุญาตให้เครื่องลูกข่าย (Client) ทำการประมวลผลคำสั่งเอสคิวแอล (SQL) ในฝั่งเซิร์ฟเวอร์ ตัวอย่างโปรแกรมเกตเวย์ เช่น ODBC (Open Database Connection) OLE-DB (Open Linking and Embedding for Database) และ JDBC (Java Database Connection) เป็นต้น
2) ชั้นกลาง (Middle Tier) เป็นโอแลปเซิร์ฟเวอร์ ที่มีการประยุกต์ใช้ โอแลปเชิงสัมพันธ์ ซึ่งเป็นส่วนขยายของ Relational DBMS โดยเชื่อมการทำงานของโครงสร้างข้อมูลหลายมิติกับการทำงานของความสัมพันธ์มาตรฐาน
3) ชั้นบน (Top Tier) เป็นเครื่องไคลเอนท์ ที่ประกอบด้วยเครื่องมือสำหรับคำถาม รายงาน วิเคราะห์ รวมถึงเครื่องมือทางดาต้าไมน์นิ่ง

โครงสร้างของคลังข้อมูล

Reed Jacobson (2000) ได้กล่าวว่าโครงสร้างของคลังข้อมูลประกอบด้วย 2 โครงสร้างที่สำคัญคือ

1) ตารางหลัก (Fact Table) เป็นตารางในคลังข้อมูลเชิงสัมพันธ์ที่เก็บค่า รายละเอียดต่างๆ ประกอบด้วยคีย์คอลัมน์ทำหน้าที่เป็นคีย์หลัก และค่าหน่วยวัด (Measure) หรือ ข้อเท็จจริง (Fact) โดยจะจัดเก็บรายละเอียดระดับล่างสุดเท่านั้น ดังตาราง 2.1

ตาราง 2.1 แสดงตัวอย่างโครงร่างแนวคิดของตารางหลักการขาย

State	Product	Month	Units	Dollars
WA	Colony Cranberry Muffins	January	3	7.95
WA	Sphinx Bagels	January	4	7.32
OR	Colony Cranberry Muffins	January	3	7.95
OR	Sphinx Bagels	January	4	7.32
WA	Colony Cranberry Muffins	February	16	42.40

ตาราง 2.2 แสดงตัวอย่างโครงร่างแท้จริงของตารางหลักการขาย

STATE_ID	PROD_ID	Month	Sales Units	Sales Dollars
1	589	1/1/1998	3	7.95
1	1218	1/1/1998	4	7.32
2	589	1/1/1998	3	7.95
2	1218	1/1/1998	4	7.32
1	589	2/1/1998	16	42.40

จากตัวอย่างโครงร่างแนวคิดของตารางหลักการขายประกอบด้วย 5 คอลัมน์ที่เก็บ State, Product, Month, Units และ Dollars ตามลำดับ โดยที่สามคอลัมน์แรกประกอบด้วย State, Product และ Month เป็นคีย์คอลัมน์ ส่วนอีกสองคอลัมน์คือ Units และ Dollars เป็นค่าหน่วยวัด ซึ่งปกติการจัดเก็บข้อมูลจริงจะมีรูปแบบดังตาราง 2.2 ที่แสดงตัวอย่างโครงร่างแท้จริงของตารางหลักการขาย คือจัดเก็บค่าคีย์เป็นตัวเลขเพื่อความรวดเร็วและลดขนาดของฐานข้อมูล

2) ตารางไดเมนชัน (Dimension Table) เป็นตารางในคลังข้อมูลเชิงสัมพันธ์ที่ประกอบด้วยแถวแต่ละแถวสำหรับสมาชิกของไดเมนชัน ซึ่งมีความเกี่ยวข้องกับลำดับขั้นความสัมพันธ์ในคลังข้อมูล คีย์คอลัมน์ของตารางไดเมนชันจะต้องประกอบด้วยค่าเฉพาะสำหรับ แต่ละสมาชิกของไดเมนชัน เรียกว่าคีย์หลัก ทำหน้าที่เชื่อมต่อกันระหว่างตารางหลักและตารางไดเมนชัน ดังตาราง 2.3

ตาราง 2.3 แสดงตัวอย่างโครงร่างแนวคิดของตารางไดเมนชันสินค้า

PROD_ID	Product Name	Subcategory
589	Colony Cranberry Muffins	Muffins
592	Colony Cranberry Muffins	Muffins
1218	Sphinx Bagels	Bagels

จากตัวอย่างโครงร่างแนวคิดของตารางไดเมนชันสินค้าประกอบด้วย 3 คอลัมน์ที่เก็บ PROD_ID, Product Name, Subcategory ตามลำดับ โดยที่คอลัมน์ PROD_ID ทำหน้าที่เป็นคีย์หลักของตารางไดเมนชันสินค้า และมีความสัมพันธ์กับโครงร่างแนวคิดแท้จริงของตารางหลักการขาย

แบบหนึ่งต่อกลุ่ม (One-to-Many Relationship) คีย์คอลัมน์นี้จะถูกเรียกว่า คีย์ภายนอก (Foreign Key)

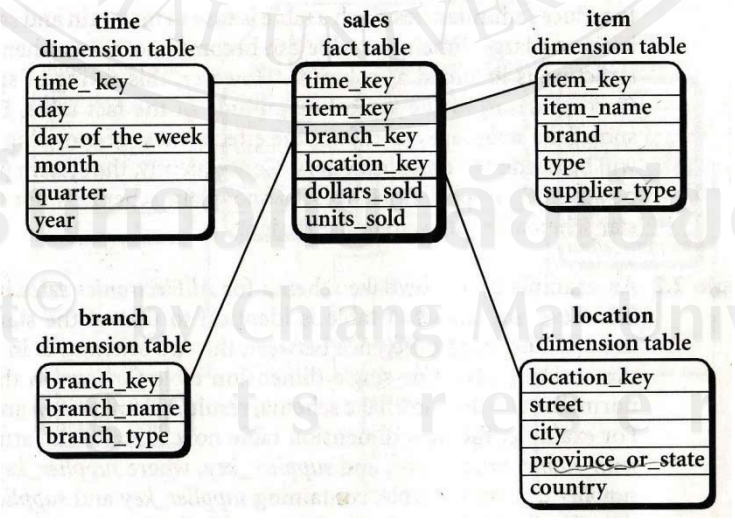
2.3 คิวบ์ข้อมูล

Han, Jiawei. and Kamber, Micheline. (2001) ได้อธิบายว่า คิวบ์ข้อมูล เป็นโมเดลที่แสดงมุมมองของข้อมูลได้ในหลายมิติ สร้างตามหลักตรรกวิทยา หรือตามเหตุและผลที่จะยอมให้แอพลิเคชันไคลเอนต์เรียกรับค่าต่างๆ ได้ถ้าค่านั้นถูกสรุปรวมไว้ในคิวแล้ว โดยที่คิวจะมีลักษณะคล้ายตารางหลัก ประกอบด้วยคอลัมน์สำหรับแต่ละคีย์หลักของไดเมนชัน คอลัมน์สำหรับค่าหน่วยวัด และการรวมค่าสมาชิกที่อาจเป็นไปได้เฉพาะสมาชิกในระดับล่างสุดของแต่ละไดเมนชันเท่านั้น นอกจากนี้คิวข้อมูลยังประกอบด้วยสมาชิก ต่างๆ จากทุกระดับของไดเมนชันด้วย

2.4 เค้าร่างของมัลติไดเมนชันนอล

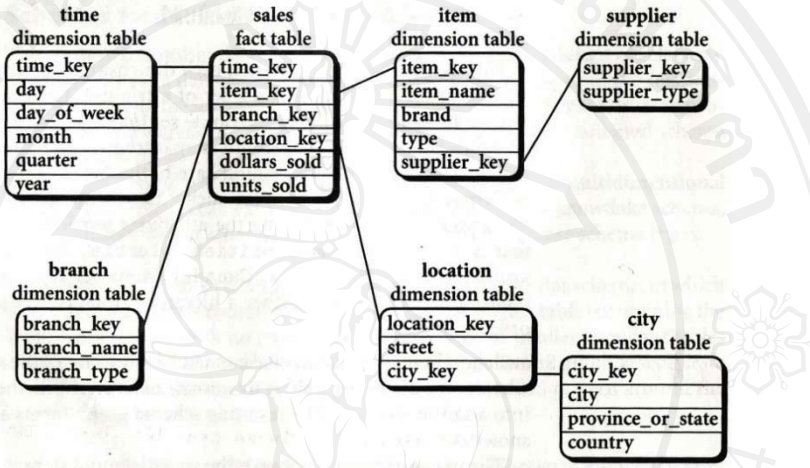
คลังข้อมูลใช้แบบจำลองมัลติไดเมนชันนอลสำหรับออกแบบความสัมพัน ธ์ของคลังข้อมูล เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมสำหรับประมวลผลโอแลป ประกอบด้วย 3 เค้าร่างดังต่อไปนี้

1) เค้าร่างดาว (Star Schema) เป็นแบบจำลองพื้นฐานของคลังข้อมูล ประกอบด้วย ตารางหลักและตารางไดเมนชัน โดยที่ตารางไดเมนชันที่เชื่อมโยงกับตารางหลักมีลักษณะคล้ายดาว ดังรูป 2.2



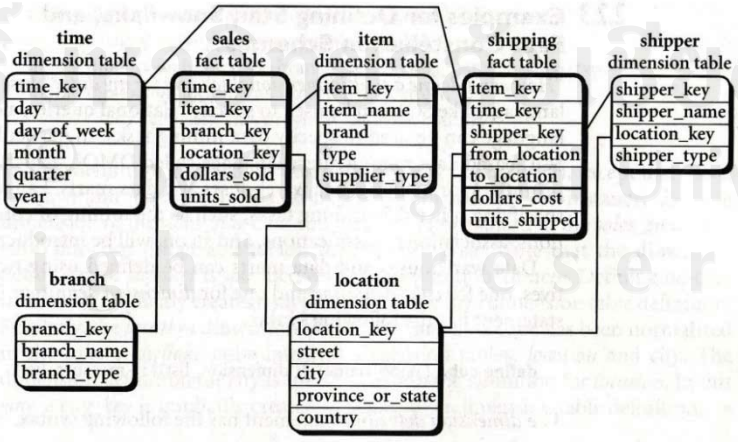
รูป 2.2 แสดงเค้าร่างดาวของคลังข้อมูลการขาย

2) เค้าร่างเกล็ดหิมะ (Snowflake Schema) เป็นแบบจำลองคลังข้อมูลที่ต่างไปจากเค้าร่างดาว เนื่องจากตารางไดเมนชันบางตารางจะถูกนอมัลไลซ์ (Normalized) เพื่อลดความซ้ำซ้อน และประหยัดพื้นที่จัดเก็บข้อมูล ทำให้มีตารางเพิ่มขึ้น แต่จำเป็นต้องมีการรวมตารางในภายหลังทำให้เสียเวลาในการประมวลผลเพิ่มขึ้น ดังรูป 2.3



รูป 2.3 แสดงเค้าร่างเกล็ดหิมะของคลังข้อมูลการขาย

3) เค้าร่างกลุ่มดาว (Fact Constellation Schema) เป็นแบบจำลองคลังข้อมูลที่มี ตารางหลักมากกว่า 1 ตาราง และมีการใช้ตารางไดเมนชันร่วมกัน ซึ่งเค้าร่างนี้อาจประกอบด้วยหลายๆ เค้าร่างดาว ดังรูป 2.4



รูป 2.4 แสดงเค้าร่างกลุ่มดาวของคลังข้อมูลการขาย และการขนส่ง