คลังข้อมูลและเทคนิคการทำเหมืองข้อมูลสำหรับการวิเคราะห์การขาย

โดย

นาย บวร น้อยแสง

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาคณิตศาสตร์และเทคโนโลยีสารสนเทศ
ภาควิชาคณิตศาสตร์
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2549

ISBN 974-11-6258-8

DATA WAREHOUSE AND DATA MINING TECHNIQUES FOR SALE ANALYSIS

By

Boworn Noisang

A Master's Report Submitted in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

Department of Mathematics

Graduate School

SILPAKORN UNIVERSITY

2006

ISBN 974-11-6258-8

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร อนุมัติให้สารนิพนธ์เรื่อง “คลังข้อมูลและเทคนิคการทำเหมืองข้อมูลสำหรับการวิเคราะห์การขาย” เสนอโดย นาย บวร น้อยแสง เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาสตรมหาบัณฑิต สาขาวิชาคณิตศาสตร์และเทคโนโลยีสารสนเทศ

.....

(รองศาสตราจารย์ ดร. ศิริชัย ชินะตังกูร)

คณบดีบัณฑิตวิทยาลัย

วันที่.....เดือน.....พ.ศ.....

ผู้ควบคุมสารนิพนธ์

ผู้ช่วยศาสตราจารย์ ดร. ปราณี นิลกรณ์

คณะกรรมการตรวจสอบสารนิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ไพบูลย์ รัตนประเสริฐ)

...../...../ .....

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. ปราณี นิลกรณ์)

...../...../ .....

.....กรรมการ

(รองศาสตราจารย์ วีรานันท์ พงศาภักดี)

...../...../ .....

K 45308306 : สาขาวิชาคณิตศาสตร์และเทคโนโลยีสารนเทศ
คำสำคัญ : คลังข้อมูล / เหมืองข้อมูล /การรวมกลุ่ม/ การวิเคราะห์การขาย

บวร น้อยแสง : คลังข้อมูลและเทคนิคการทำเหมืองข้อมูลสำหรับการวิเคราะห์การขาย (DATA WAREHOUSE AND DATA MINING TECHNIQUES FOR SALE ANALYSIS) อาจารย์ผู้ควบคุมสารนิพนธ์ :ผศ. ดร. ปราณี นิลกรณ์ .80 หน้า .ISBN 974-11-6258-8

การศึกษาในสารนิพนธ์ฉบับนี้ มีวัตถุประสงค์เพื่อพัฒนาคลังข้อมูลของระบบการขาย และประยุกต์เทคนิคการทำเหมืองข้อมูลกับคลังข้อมูลที่พัฒนาขึ้นสำหรับช่วยในการวิเคราะห์การขาย ซึ่งในการทำเหมืองข้อมูลจะเป็นการจัดกลุ่มลูกค้าโดยใช้วิธีแบบ K-Means ใช้ระยะทางแบบยุคลิด ในการรวมกลุ่มลูกค้า ข้อมูลที่ใช้ในการพัฒนาและวิเคราะห์ได้มาจากระบบการขายของบริษัทแฟนซีอาร์ท จำกัดตั้งแต่เดือนตุลาคม พ.ศ. 2547 ถึง เดือนกรกฎาคม พ.ศ. 2548 มีผลการพัฒนาแบ่งเป็น 2 ส่วนคือ

ส่วนการพัฒนาคลังข้อมูลใช้รูปแบบ Schema คือ Star Schema ประกอบด้วยตัววัดคือ จำนวนรวมของสินค้าที่ขาย ราคาสินค้าเฉลี่ย ต้นทุนเฉลี่ยของสินค้า มูลค่ารวมในการขายสินค้า และมีมิติ(Dimension)อยู่ 4 มิติคือ เวลา สินค้า ลูกค้า พนักงานขาย

ส่วนการทำเหมืองข้อมูล ได้นำข้อมูลลูกค้าที่ได้จากคลังข้อมูลซึ่งมีจำนวน 760 ราย มาวิเคราะห์การรวมกลุ่มแบบ K-Means ตัวแปรที่ใช้ในการวิเคราะห์มี 6 ตัวแปรคือ จำนวนรวมสินค้าทั้งหมดที่ขายสินค้า ราคาค่าเฉลี่ยของสินค้า ต้นทุนเฉลี่ย มูลค่ารวมที่ขายสินค้าทั้งหมด เกรดของบริษัทลูกค้าที่กำหนดและวงเงินสินเชื่อในการขายสินค้าของบริษัทลูกค้า สามารถรวมกลุ่มลูกค้าได้ 5 กลุ่ม คือ

กลุ่มที่ 1 มีจำนวนลูกค้า 41 ราย เป็นกลุ่มที่มีวงเงินสินเชื่อในการขายสินค้ามูลค่ามากที่สุด และเป็นกลุ่มที่ทำรายได้ให้กับบริษัทมากที่สุด

กลุ่มที่ 2 มีจำนวนลูกค้า 2 ราย เป็นบริษัทสาขาที่เปิดเป็นร้านขายปลีกจากทั้งหมด 4 บริษัท เป็นกลุ่มที่ทำรายได้ให้กับบริษัทน้อยที่สุด

กลุ่มที่ 3 มีจำนวนลูกค้า 2 ราย เป็นบริษัทสาขาที่เปิดเป็นร้านขายปลีกที่เหลืออีก 2 ราย แต่มีมูลค่ารวมการขายสินค้ามากกว่ากลุ่มที่ 2

กลุ่มที่ 4 มีจำนวนลูกค้า 48 ราย เป็นกลุ่มที่ซื้อสินค้าที่มีราคาเฉลี่ยสูงสุด และเป็นกลุ่มที่ทำรายได้ให้กับบริษัทเป็นอันดับที่สอง

กลุ่มที่ 5 มีจำนวนลูกค้า 667 ราย เป็นกลุ่มที่มีขนาดใหญ่ที่สุด มีค่าเฉลี่ยของมูลค่ารวมของการขายสินค้าน้อยที่สุด แต่ก็เป็นกลุ่มที่ทำรายได้ให้กับบริษัทเป็นอันดับสาม

ภาควิชาคณิตศาสตร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ปีการศึกษา 2549

ลายมือชื่อนักศึกษา .....

ลายมือชื่ออาจารย์ผู้ควบคุมสารนิพนธ์ .....

K 45308306 : MAJOR : MATHEMATICS AND INFORMATION TECHNOLOGY
KEY WORD : DATA WAREHOUSE / DATA MINING/CLUSTERING/SALE ANALYSIS
BOWORN NOISANG : DATA WAREHOUSE AND DATA MINING TECHNIQUES FOR SALE ANALYSIS . MASTER'S REPORT ADVISOR : ASST. PROF. PRANEE NILAKORN , Ph.D. 80 pp. ISBN 974-11-6258-8

The objective of this project were to develop a data warehouse for sales system and to apply data mining techniques to obtain information useful for sale analysis from the data warehouse. The data mining technique used in this study was cluster analysis with K-Means algorithm based on Euclidean distance. The data used in the development and analysis were from sales system of Fancy Art company between October , 2004 to July , 2005 .

A star schema was used with the developed data warehouse with the following measures: total quantities of products sold, average price, average cost and total value sale and under the following dimensions: time, product, customer and sale employee.

The total of 670 customers from the data warehouse were analyzed by means of cluster analysis with total quantity of products sold, average price, average cost, total value sale of product, grade of customer's company and financial credit of customer's company as classifying variables. According to our study, customers should be grouped in to 5 clusters. The profiles of each cluster are as follows :

Cluster 1 contains 41 customers, contributes the most to the income of Fancy Art company and obtains most of the financial credit from Fancy Art.

Cluster 2 contains 2 customers which are retail branching companies, contributes the least to the income of Fancy Art.

Cluster 3 also contains 2 customers which are also retailers, but contributes more to the income of Fancy Art than those in cluster 2.

Cluster 4 contains 48 customers, ranked second in the contribution to the income of Fancy Art. This cluster buys products from Fancy Art with highest average price.

Cluster 5, the largest cluster, contains 667 customers, ranked third in the contribution to the income of Fancy Art. The average of total value sale of this group is least.

Department of Mathematics Graduate School, Silpakorn University Academic Year 2006

Student's signature ……..

Master's Report Advisor's signature ……..

กิตติกรรมประกาศ

ในการศึกษาและเรียบเรียงสารนิพนธ์ฉบับนี้ สามารถเสร็จสมบูรณ์ไปได้ดี ก็ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร. ปราณี นิลกรณ์ อาจารย์ผู้ควบคุมสารนิพนธ์ ที่คอยดูแลเอาใจใส่ให้คำแนะนำ ให้คำปรึกษา และช่วยแก้ไขข้อบกพร่องต่างๆในสารนิพนธ์ฉบับนี้

ขอบกราบขอบพระคุณคณะอาจารย์ภาควิชาคณิตศาสตร์ ภาควิชาสถิติ และภาควิชาคอมพิวเตอร์ มหาวิทยาลัยศิลปากรทุกท่านที่ประสิทธิ์ประสาท วิชา ความรู้ ต่างๆ และคอยดูแลเอาใจใส่และให้ความช่วยเหลือด้วยดีตลอดมา

ขอขอบคุณ คุณ สุขุมา ปิยะสวัสดิ์ ผู้จัดการแผนกสารสนเทศ บริษัทแฟนซีอารท์ จำกัด ที่คอยช่วยเหลือในเรื่องต่างๆ ในการทำสารนิพนธ์ฉบับนี้

สุดท้ายนี้ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ ที่คอยสนับสนุนและเป็นกำลังใจเสมอมา

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในการประกอบธุรกิจขององค์กรธุรกิจใดๆย่อมมีวัตถุประสงค์ที่จะทำให้องค์กรได้รับผลประโยชน์สูงสุด ภายใต้การแข่งขันกันกับองค์กรอื่นๆ ปัจจัยหนึ่งที่ทำให้องค์กรประสบความสำเร็จ คือ องค์กรจำเป็นต้องมีการเก็บบันทึกข้อมูลเกี่ยวกับกิจกรรมหรือการดำเนินการขององค์กรไว้ ปัจจุบันคอมพิวเตอร์มีบทบาทอย่างมากในการนำมาช่วยงานทางด้านธุรกิจ และเครื่องมือที่สำคัญที่ช่วยในการวิเคราะห์ข้อมูล ซึ่งเราเรียกระบบคอมพิวเตอร์ที่ใช้เพื่อการนี้ว่า ระบบสารสนเทศ (Information system) หน้าที่หลักของระบบสารสนเทศคือการนำข้อมูล (Data) ที่มีอยู่ในองค์กรมาประมวลผล เพื่อให้เกิดสารสนเทศ(Information)ที่สามารถนำไปใช้ประโยชน์ได้ในการบริหารองค์กร (กิตติพงศ์ กลมกล่อม 2548 : 2)

การขายหรือการจัดจำหน่ายก็เป็นกิจกรรมที่สำคัญอย่างหนึ่งขององค์กรธุรกิจ เนื่องจากการขายเป็นการนำรายได้เข้ามาล่อเลี้ยงองค์กรให้สามารถดำเนินกิจการต่อไปได้ ดังนั้นการวิเคราะห์การขายจึงเป็นเครื่องมือชนิดหนึ่ง ที่ใช้ในการบริหารขององค์กร ในการวิเคราะห์การขายโดยทั่วไปนั้นจะเป็นการนำเสนอในรูปแบบรายงาน เช่นรายงานสินค้าที่ขายสูงสุดหรือตามลำดับจากมากไปน้อย รายงานสินค้าที่ไม่เคลื่อนไหว รายงานยอดขายของพนักงานขาย (ศรัณย์ ชูเกียรติ 2547:706)เป็นต้น ซึ่งรายงานส่วนใหญ่จะรายงานเพื่อช่วยในการตัดสินใจของผู้บริหาร ข้อมูลดังกล่าวนั้นต้องถูกต้องและทันต่อเหตุการณ์ที่เปลี่ยนแปลงตลอดเวลา

ปัจจุบันข้อมูลที่เกิดจากการทำธุรกรรม (Transaction)มีจำนวนมากมาย การนำข้อมูลดังกล่าวมาใช้ได้อย่างเหมาะสมและเป็นประโยชน์ต่อองค์กรมีความสำคัญเป็นอย่างมากโดยเฉพาะอย่างยิ่งในสภาพการแข่งขันที่มีความรุนแรงกันมากขึ้น วิธีที่นิยมของผู้บริหารแบบนี้ของบางองค์กรในการนำข้อมูลจำนวนมากที่มีอยู่มาใช้คือ การทำระบบคลังข้อมูล การทำคลังข้อมูล(Data warehousing)จะเป็นการช่วยการทำการเก็บรวบรวมข้อมูลจำนวนมากและสามารถวิเคราะห์แบบหลายมิติ(Multidimensional data analysis)ได้ หรือการทำการประมวลผลในเชิงวิเคราะห์แบบออนไลน์(On-Line-Analytic Processing หรือ OLAP) อีกทั้งยังสามารถเรียกใช้ข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพ เพื่อเป็นประโยชน์ต่อการตัดสินใจของผู้บริหาร การทำคลังข้อมูลนั้นยัง

ลดปัญหาที่เกิดขึ้นจากการใช้ข้อมูลจากฐานข้อมูลปฏิบัติงาน(Operational database) โดยปัญหาที่พบจากการเก็บข้อมูลในรูปของการทำธุรกรรม(Transaction system)เช่น

- การเรียงข้อมูลจากฐานข้อมูลขนาดใหญ่ จะทำได้ช้าและทำให้ประสิทธิภาพของระบบลดลง
- ไม่มีความยืดหยุ่นหรือไม่สามารถเปลี่ยนแปลงรูปแบบของสารสนเทศตามความต้องการผู้ใช้งานได้
- ไม่สามารถวิเคราะห์ข้อมูลที่ซับซ้อน เช่น การพยากรณ์
- ไม่มีการเก็บข้อมูลย้อนหลัง
- ไม่ตอบสนองการทำคิวรี่(Query) ที่ซับซ้อนได้
- ข้อมูลถูกจัดเก็บอยู่ตามฐานข้อมูลต่าง ๆ หลายฐานข้อมูลทำให้ยากต่อการเข้าถึงและเรียกใช้

การทำเหมืองข้อมูล(Data mining) ก็เป็นอีกวิธีการหนึ่งที่เป็นการวิเคราะห์ข้อมูลในฐานข้อมูลเพื่อช่วยในการตัดสินใจของผู้บริหาร ซึ่งเหมืองข้อมูล จะเป็นการค้นหา วิเคราะห์ หรือสร้างองค์ความรู้ใหม่จากข้อมูลขนาดใหญ่ซึ่งอาจจะเป็นการค้นหารูปแบบหรือกฎ โดยการใช้เทคนิคทางสถิติ ทางคณิตศาสตร์หรือเทคนิคทางวิทยาการคอมพิวเตอร์ ในบางครั้งการทำเหมืองข้อมูลนั้นอาจถูกกล่าวถึงใน Knowledge discovery หรือ KDD(Knowledge Discovery in Database)เนื่องจากเหมืองข้อมูลเป็นขั้นตอนที่สำคัญใน KDD การรวมกลุ่ม(Clustering)ก็เป็นวิธีการอย่างหนึ่งที่ใช้ในการทำเหมืองข้อมูล โดยการรวมกลุ่มมีหลายเทคนิคด้วยกัน เช่นเทคนิครวมกลุ่มแบบมีลำดับชั้น (Hierarchical clustering technique) และเทคนิค K-Means

การวิเคราะห์การขายมีประโยชน์ในการตัดสินใจของผู้บริหารเกี่ยวกับนโยบายส่งเสริมการขายต่าง ๆ จึงจำเป็นต้องทำอย่างรวดเร็ว ถูกต้อง และทันต่อเหตุการณ์ คลังข้อมูลและเทคนิคการทำเหมืองข้อมูลจะช่วยทำให้การทำการวิเคราะห์การขายมีความรวดเร็วและมีประสิทธิภาพมากขึ้น อีกทั้งยังสามารถดึงสารสนเทศที่ซ่อนอยู่ในข้อมูลออกมา เพื่อช่วยในการตัดสินใจในการบริหารงานให้ทันต่อการเปลี่ยนแปลงในทางธุรกิจ เช่น การรวมกลุ่มลูกค้าที่มีลักษณะคล้ายคลึงกัน จะสามารถทำให้เรากำหนดกิจกรรมต่างๆที่จะกระทำต่อลูกค้ามีประสิทธิภาพ ถูกกลุ่มเป้าหมายเพิ่มมากขึ้น เพื่อเพิ่มยอดขายให้กับองค์กร ในงานวิจัยนี้จึงสนใจศึกษาการนำคลังข้อมูลและเทคนิคการทำเหมืองข้อมูลมาใช้ในการวิเคราะห์การขาย

วัตถุประสงค์ของการศึกษา

1. พัฒนาคลังข้อมูลของระบบขาย โดยใช้กรณีศึกษาการขายของบริษัทแฟนซีอารท์ จำกัด และบริษัทในเครือ ซึ่งประกอบธุรกิจผลิตและจัดจำหน่ายเสื้อผ้าสำเร็จรูปและของชำร่วย
2. ประยุกต์เทคนิคการทำเหมืองข้อมูลกับคลังข้อมูลที่พัฒนาขึ้น เพื่อช่วยในการวิเคราะห์

การขาย โดยใช้วิธีการรวมกลุ่ม(Clustering) เพื่อรวมกลุ่มลูกค้าที่คล้ายคลึงสำหรับช่วยในการทำการส่งเสริมการขายหรือการสร้างความสัมพันธ์กับลูกค้า

ขอบเขตของการศึกษา

ในการศึกษานี้จะใช้ข้อมูลการขายของบริษัทแฟนซีอารท์ ตั้งแต่เดือนตุลาคม พ.ศ. 2547 ถึงเดือนกรกฎาคม พ.ศ. 2548 เพื่อใช้เป็นข้อมูลต้นแบบในการพัฒนาระบบดังต่อไปนี้

1. สร้างคลังข้อมูลสำหรับวิเคราะห์การขาย พร้อมทั้งนำข้อมูลมาวิเคราะห์แบบหลายมิติ (Multidimensional data analysis) หรือการประมวลผลในเชิงวิเคราะห์แบบออนไลน์(On-Line Analytic Processing หรือ OLAP)

2. ใช้การวิเคราะห์การรวมกลุ่ม(Cluster analysis) แบบ K-Means เป็นกรณีศึกษาในการทำเหมืองข้อมูลเพื่อรวมกลุ่มลูกค้าจากข้อมูลการขาย

ประโยชน์ที่คาดว่าจะได้รับ

1. การสร้างคลังข้อมูลและเหมืองข้อมูลสำหรับการวิเคราะห์การขาย จะสามารถช่วยให้ผู้บริหารมีสารสนเทศที่รวดเร็วและมีประสิทธิภาพ เพื่อสนับสนุนการตัดสินใจในการดำเนินธุรกิจ

2. เป็นกรณีศึกษาในการทำเหมืองข้อมูล โดยใช้วิธีการทางสถิติเข้ามาช่วยทางด้านงานธุรกิจ เพื่อเป็นตัวอย่างสำหรับการนำเอาวิธีทางสถิติอื่นๆ มาใช้ในทางธุรกิจ

3. เพื่อช่วยให้องค์กรมีประสิทธิภาพด้านการขายโดยสามารถทำการส่งเสริมการขายหรือการสร้างความสัมพันธ์กับลูกค้าให้ถูกกลุ่มลูกค้ามากขึ้น

คำจำกัดความที่ใช้ในการศึกษา

1. ลูกค้า หมายถึง บริษัท ห้างร้าน กลุ่มบุคคล บุคคล ที่ซื้อสินค้ากับบริษัทที่มีข้อมูลอยู่ในฐานข้อมูล

2. ต้นทุนของสินค้า หมายถึง ต้นทุนที่ใช้การผลิตสินค้าแต่ละชิ้นที่บริษัทผลิตขึ้น

3. เกรดของลูกค้า หมายถึง ระดับของลูกค้าโดยพิจารณาจากปัจจัยต่าง ๆ เช่น ฐานะทางการเงิน ระยะเวลาที่ติดต่อกับบริษัท มูลค่าซื้อขาย เป็นต้น

4. วงเงินสินเชื่อในการขายสินค้าของลูกค้า หมายถึง วงเงินสินเชื่อที่บริษัทกำหนดให้กับลูกค้าในการซื้อสินค้า ในกรณีที่ลูกค้าซื้อสินค้าโดยใช้วงเงินสินเชื่อ

บทที่ 2

วรรณกรรมที่เกี่ยวข้อง

วรรณกรรมที่เกี่ยวข้องที่จะนำเสนอในบท จะนำเสนอเรียงตามลำดับต่อไปนี้

1. การขายและการวิเคราะห์การขาย
2. คลังข้อมูลและการคลังข้อมูล
3. การทำเหมืองข้อมูล
4. งานวิจัยที่เกี่ยวข้อง
5. บริษัทแฟนซีอารท์

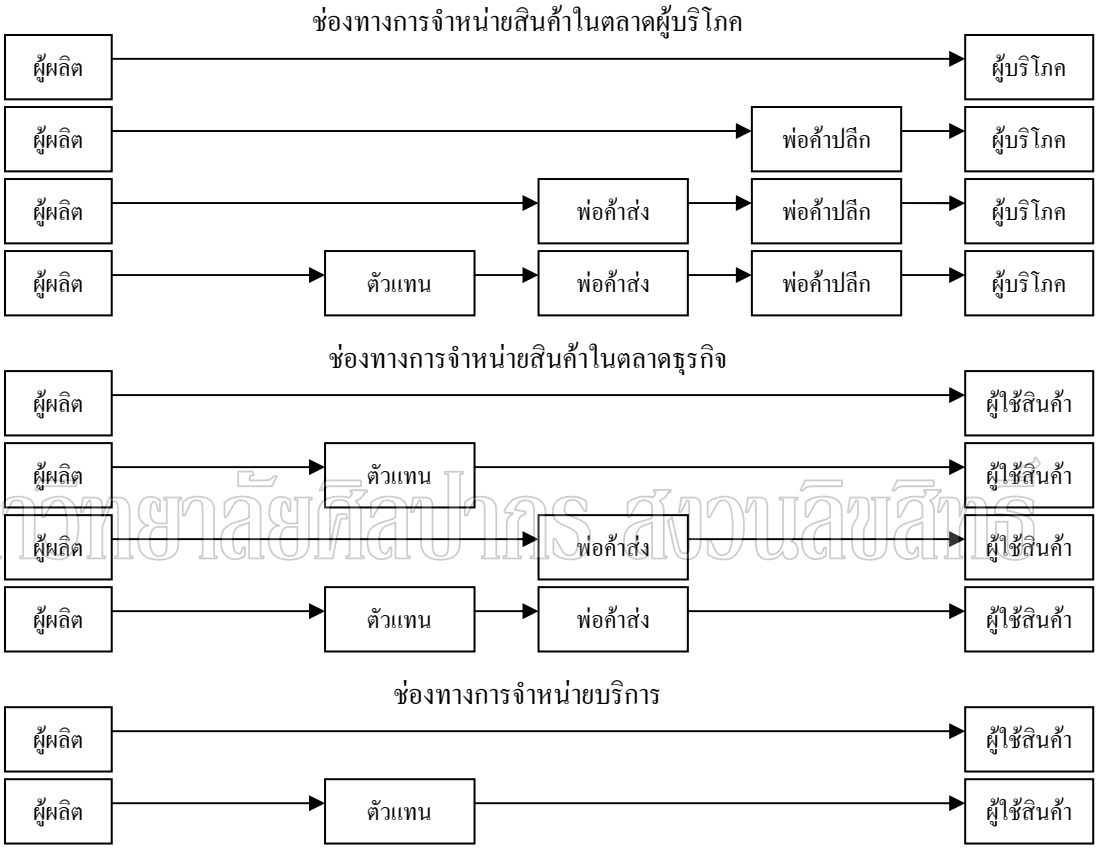
1. การขายและการวิเคราะห์การขาย (Sale and sale analysis)

การขายตามประมวลกฎหมายแพ่งและพาณิชย์ มาตรา 453 กล่าวว่า “อันว่าการซื้อนั้น คือสัญญาซึ่งบุคคลฝ่ายหนึ่งเรียกว่าผู้ขาย โอนกรรมสิทธิ์แห่งทรัพย์สินให้แก่บุคคลอีกฝ่ายหนึ่ง เรียกว่าผู้ซื้อและผู้ซื้อตกลงว่าจะใช้ราคาแห่งทรัพย์สินนั้นให้แก่ผู้ขาย ”(วารินทร์ สินสูงสุด 2539 : 37) นอกจากนี้ยังมีความหมายหรือคำจำกัดความของการขายอื่น ๆ อีก เช่น “ การขาย (Selling) เป็นกระบวนการชักจูงผู้มุ่งหวังแบบเฉพาะบุคคล หรือไม่เฉพาะบุคคลให้ซื้อสินค้าหรือบริการ หรือนิยมชมชอบความคิดของผู้ขาย ซึ่งผู้ขายก็ได้รับผลประโยชน์ทางการค้า ” หรืออีกความหมายหนึ่ง “เป็นการใช้ศิลปะการเป็นผู้นำเพื่อชักจูงคนให้ซื้อสินค้าและบริการ” (วารินทร์ สินสูงสุด 2539 : 39)

การขายหรือจัดจำหน่าย(Distribution)มีความสำคัญต่อองค์กร เนื่องจากเป็นกิจกรรมที่ทำให้องค์กรสามารถประกอบธุรกิจต่อไปได้ ซึ่งการนำสินค้าหรือบริการไปถึงผู้บริโภคนั้นมีหลายวิธีการ ซึ่งวิธีการเหล่านั้นเราจะเรียกว่าช่องทางจัดจำหน่าย(Channel of distribution หรือ Distribution channel หรือ Marketing channel) ดังนั้น ช่องทางการจัดจำหน่าย หมายถึง กลุ่มคนและกิจกรรมที่เกี่ยวข้องกับการเคลื่อนย้ายกรรมสิทธิ์ในสินค้าหรือบริการจากผู้ผลิตไปยังผู้บริโภคขั้นสุดท้าย(Ultimate consumer) หรือผู้ใช้ทางธุรกิจ(Business user) หรือผู้ใช้ทางอุตสาหกรรม (Industrial users) (ศิริวรรณ เสรีรัตน์ 2543 : 143)ซึ่งจะแบ่งออกเป็น 2 ลักษณะคือ

1. ช่องทางการจัดจำหน่ายทางตรง(Direct channel) หรือการขายตรง(Direct selling) หรือตลาดทางตรง(Direct marketing) หรือ การจัดจำหน่ายทางตรง(Direct distribution) หมายถึง การขายสินค้าจากผู้ผลิตไปยังผู้บริโภคหรือผู้ใช้ทางอุตสาหกรรมโดยไม่มีคนกลาง

2. ช่องทางการจัดจำหน่ายทางอ้อม(Indirect channel หรือ Indirect distribution) หมายถึง เส้นทางที่สินค้าเคลื่อนย้ายจากผู้ผลิตไปยังลูกค้าโดยต้องผ่านคนกลาง



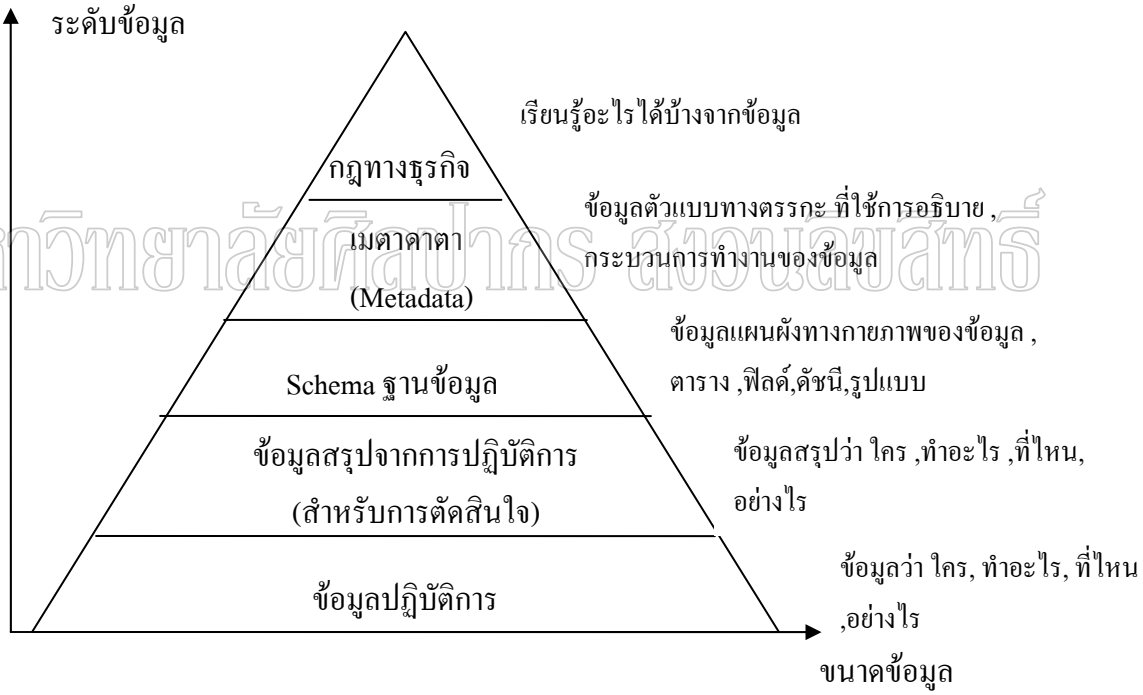
ภาพที่ 1 ช่องทางการจำหน่าย
ที่มา : สุดาดวง เรืองรุจิระ , หลักการตลาด (กรุงเทพฯ : สำนักพิมพ์ประกายพรึก,2543),207.

การวิเคราะห์การขายจึงเป็นการวิเคราะห์ข้อมูลที่ได้มาจากกิจกรรมของการขายหรือการจัดจำหน่าย ระบบสารสนเทศสำเร็จรูปที่ใช้ในการวิเคราะห์การขายจึงเป็นการสรุปข้อมูล หรือกราฟเป็นส่วนใหญ่ นอกเสียจากเป็นระบบสารสนเทศเพื่อเป็นการวิเคราะห์โดยเฉพาะจึงจะมีการวิเคราะห์ที่ซับซ้อนมากขึ้น ตัวอย่างของการวิเคราะห์การขายเช่น การประมาณการขาย ข้อมูลสรุปการขายตามช่วงเวลาต่างๆ การเปรียบเทียบข้อมูลการขาย การขายสินค้าสูงสุดและต่ำสุด ซึ่งข้อมูล

เหล่านี้จะช่วยในการตัดสินใจและการวางแผนบริหารของผู้บริหารให้มีประสิทธิ์ภาพมากขึ้น เพื่อเป็นการเพิ่มยอดขายให้มากขึ้นซึ่งผลที่ได้ตามมาคือผลกำไรของการประกอบการที่มากขึ้น

2. คลังข้อมูลและการคลังข้อมูล (Data warehouse and Data warehousing)

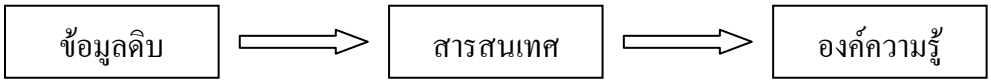
การนำคอมพิวเตอร์เข้ามาช่วยในการดำเนินการข้อมูลในทางธุรกิจมีมาตั้งแต่ปี ค.ศ. 1960 และในกลางปี ค.ศ. 1980 เริ่มนิยมนำฐานข้อมูลเชิงสัมพันธ์(Relational database)มาใช้ในทางธุรกิจ ทำให้มีการสร้างระบบปฏิบัติงานหรือระบบ OLTP (On-Line Transaction Processing)ในธุรกิจโดยคอมพิวเตอร์ส่งผลให้มีการเก็บข้อมูลอย่างต่อเนื่อง ทำให้จำนวนข้อมูลมีขนาดใหญ่ ข้อมูลเหล่านี้เราจะเรียกว่าข้อมูลการปฏิบัติงาน(Operational data) หรือข้อมูลธุรกรรม(Transaction data)



ภาพที่ 2 ความสัมพันธ์ของขนาดข้อมูลกับระดับข้อมูล

ที่มา : Michael J.A. Berry and Gordon S. Linoff , Data mining Techniques :For Marketing, Sales,and Customer Relationship Management (Indiana : Wiley Publishing Inc, 2004),475.

จากจำนวนและความซ้ำซ้อนของข้อมูลที่มีมากจากข้อมูลปฏิบัติงานทำให้การได้มาของสารสนเทศนั้นไม่ทันต่อเหตุการณ์(Not the right information at the right time)(Berry And Linoff 2004 : 473) คลังข้อมูลจะช่วยทำให้การได้มาของสารสนเทศทันต่อเหตุการณ์



ภาพที่ 3 กระบวนการสร้างองค์ความรู้

คลังข้อมูล(Data warehouse) เป็นการเก็บข้อมูลในเชิงหัวข้อ(Subject-oriented)ซึ่งทำการรวบรวมข้อมูลมาจากแหล่งต่างๆ(Integrated)โดยจะเก็บข้อมูลเป็นระยะเวลานาน(Time-variant) และ ไม่มีการเปลี่ยนแปลงของชุดข้อมูลได้โดยง่าย(Non-volatile)ซึ่งเป็นข้อมูลสำหรับที่ใช้ช่วยการสนับสนุนในการดำเนินการตัดสินใจเพื่อการบริหาร (Roiger and Geatz 2003 : 184)

- Subject-oriented คือข้อมูลจะถูกสร้างและถูกเก็บเก็บในเชิงหัวข้อโดยจะเก็บเฉพาะข้อมูลที่จำเป็นต่อกระบวนการตัดสินใจ

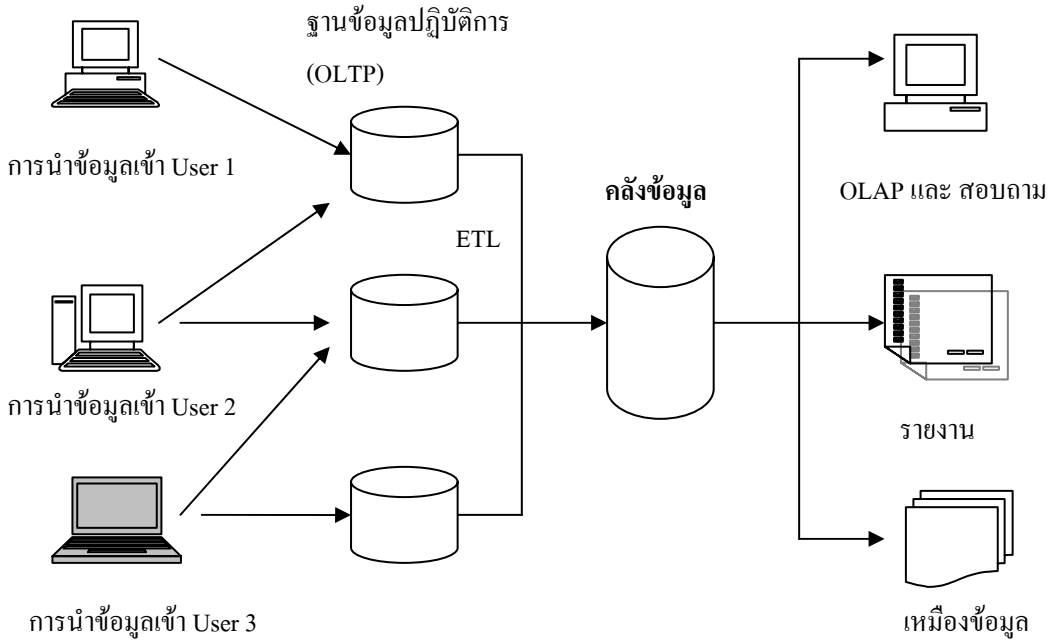
- การรวบรวม (Integrated)เป็นการรวบรวมข้อมูลจากแหล่งต่าง ๆ ทั้งภายในและภายนอกองค์กรที่เกี่ยวกับเซตข้อมูลที่เราต้องการให้มาอยู่ที่เดียวกัน(ในฐานข้อมูลเดียวกัน)พร้อมทั้งทำให้ข้อมูลที่มาจากต่างแหล่งกันนั้นมีความสอดคล้อง

- Time variant ในการเก็บข้อมูลในคลังข้อมูลนั้นจะใช้ระยะเก็บข้อมูลนานเพื่อช่วยการวิเคราะห์ข้อมูล

- Non-volatile ข้อมูลที่อยู่ในคลังข้อมูลจะไม่มีการเปลี่ยนแปลงหรือแก้ไขง่ายๆ ผู้ใช้สามารถโหลดหรือเข้าถึงข้อมูลได้เท่านั้น (สุนีย์ พงษ์พินิจภิญโญ ม.ป.ป. : 418)

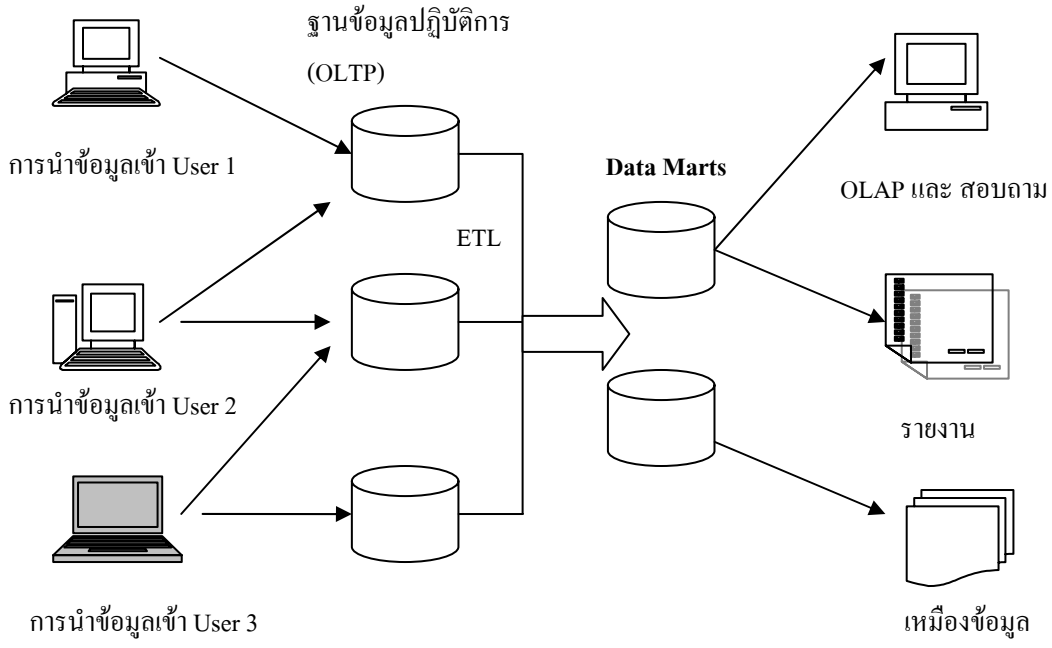
การคลังข้อมูล(Data warehousing)คือการออกแบบและสร้างโครงสร้างของข้อมูลในคลังข้อมูล การเพื่อให้ได้มาซึ่งข้อมูล การสร้างผลลัพธ์จากข้อมูลที่มี รวมไปถึงการรักษา การปรับปรุงประสิทธิภาพ รวมทั้งวิธีการต่างๆ ที่เกี่ยวข้องกับคลังข้อมูล

ในวิธีการเพื่อให้ได้มาซึ่งข้อมูลนั้นบางครั้งต้องเข้าถึง(Access)ข้อมูลที่มาจากต่างแหล่งข้อมูล(Data source)เพื่อนำมารวมเข้าไว้ด้วยกัน จะต้องมีการทำปรับข้อมูลเพื่อลดความซ้ำซ้อน ความสอดคล้องของข้อมูลที่มาจากต่างแหล่งกัน และความผิดพลาดของข้อมูล รวมทั้งการเลือกข้อมูลที่เป็นประโยชน์(Filtering)ซึ่งกระบวนการETL(Extraction,Transformation ,and Load)จะช่วยในการดำเนินการในการโอนถ่ายข้อมูล โดยข้อมูลจะถูกจัดเก็บในลงในฐานข้อมูลที่จะเรียกว่า Data warehouse database บางครั้งเราอาจสร้างชุดข้อมูลที่มีขนาดเล็กกว่าแต่มีลักษณะเหมือนกับคลังข้อมูลเพื่อช่วยการวิเคราะห์ข้อมูลบางปัญหาเฉพาะปัญหา ซึ่งชุดข้อมูลที่สร้างขึ้นเราเรียกว่า Data marts รูปดังต่อไปนี้จะแสดงถึงสถาปัตยกรรมต่าง ๆ ของคลังข้อมูล



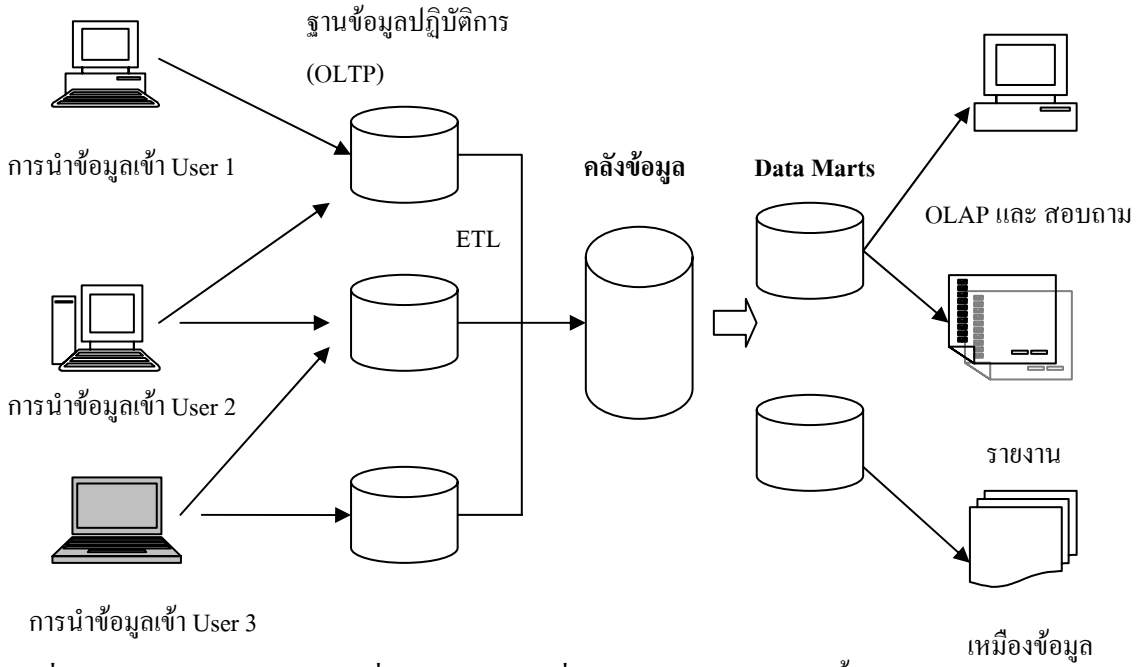
ภาพที่ 4 สถาปัตยกรรมคลังข้อมูลที่ไม่มี Data marts

ที่มา : Sakhr Youness , Professional Data Warehousing with SQL Server 7.0 and OLAP Services (Birmingham : Wrox Press , 2005), 19 .



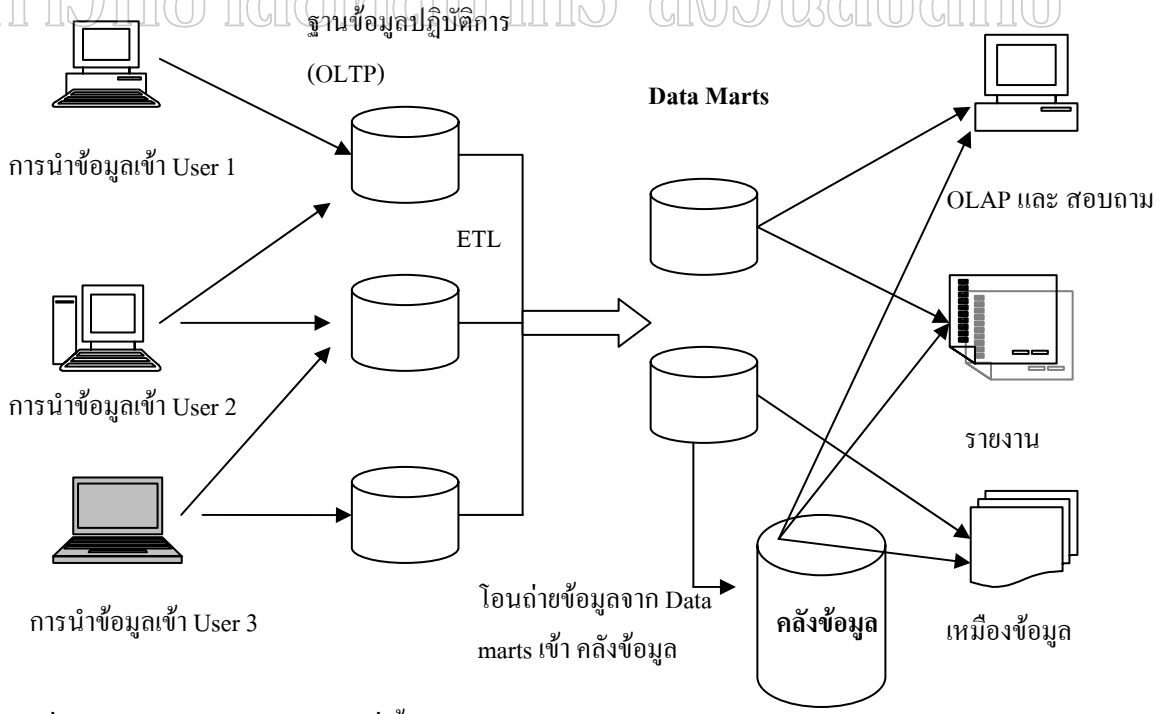
ภาพที่ 5 สถาปัตยกรรมคลังข้อมูลที่ใช้ Data marts แทน คลังข้อมูล

ที่มา : Sakhr Youness , Professional Data Warehousing with SQL Server 7.0 and OLAP Services (Birmingham : Wrox Press , 2005), 19 .



ภาพที่ 6 สถาปัตยกรรมคลังข้อมูลที่ใช้ Data marts ที่สร้างจากคลังข้อมูลเท่านั้น

ที่มา : Sakhr Youness , Professional Data Warehousing with SQL Server 7.0 and OLAP Services (Birmingham : Wrox Press , 2005), 20 .



ภาพที่ 7 สถาปัตยกรรมคลังข้อมูลที่ทั้งคลังข้อมูลและ Data marts

ที่มา : Sakhr Youness , Professional Data Warehousing with SQL Server 7.0 and OLAP Services (Birmingham : Wrox Press , 2005), 20 .

2.1 แบบจำลองข้อมูลสำหรับคลังข้อมูล

แบบจำลองข้อมูลเป็นเอกสารที่ใช้แสดงถึงโครงสร้างของข้อมูลว่าข้อมูลมีอิสระหรือมีความสำพันธ์กับข้อมูลอื่นอย่างไร โดยทั่วไปแล้วเทคนิคแบบจำลองข้อมูลที่ใช้ในการออกแบบคลังข้อมูลมีอยู่ 2 เทคนิคคือ ตัวแบบเชิงสัมพันธ์(Relational modeling)และ ตัวแบบเชิงมิติ (Dimensional modeling)

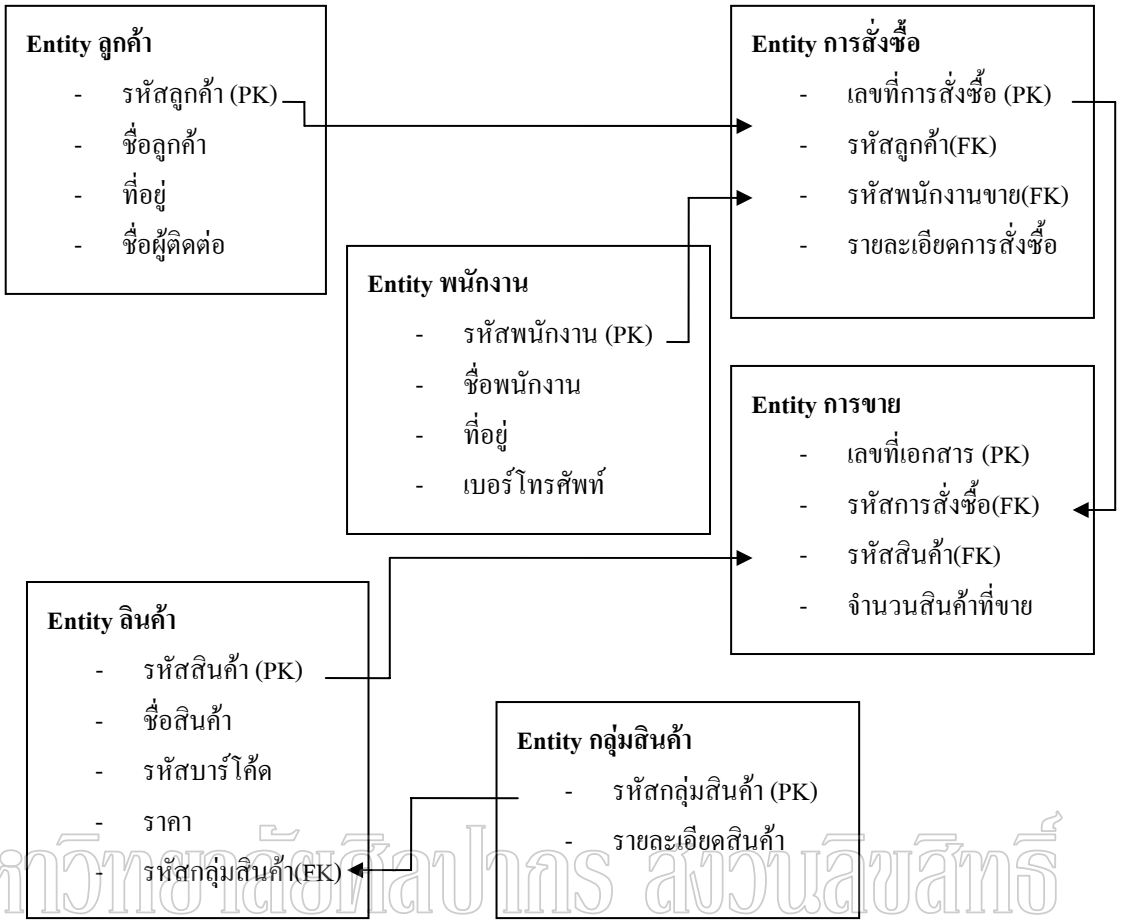
2.1.1 ตัวแบบเชิงสัมพันธ์(Relational Modeling)

คำที่จะได้พบเมื่อกล่าวถึงตัวแบบข้อมูลเชิงสัมพันธ์คือ ER Diagram (Entity Relationship Diagram หรือ ERD) ER Diagram เป็นเครื่องมือที่จะใช้แสดงโครงสร้างของข้อมูลในเทอมของ entity และความสัมพันธ์ของ entity โดย entity จะแสดงถึงวัตถุ(Object)ที่สามารถสังเกตและจำแนกโดยคุณสมบัติและคุณลักษณะ โดยแอททริบิวท์(Attributes)จะอธิบายถึงคุณลักษณะหรือคุณสมบัติของ entity กล่าวคือ แอททริบิวท์ถูกกำหนดเป็นส่วนประกอบของentity เช่น entity ของพนักงานจะมีแอททริบิวท์ที่อธิบายคุณลักษณะคือ ชื่อ ที่อยู่ ตำแหน่ง เงินเดือน เป็นต้น ความสัมพันธ์ระหว่าง entity โดยจะมีแอททริบิวท์ที่มีคุณสมบัติมีความเป็นเอกลักษณ์(Uniqueness property) เป็นสิ่งที่กำหนดความเป็นเอกลักษณ์ของแต่ละแถวซึ่งจะเรียกว่าคีย์(Key) และยังเป็นตัวเชื่อมโยงความสัมพันธ์ระหว่าง entity ซึ่งคีย์มีหลายประเภทเช่น คีย์หลัก(Primary key),คีย์รอง (Secondary key) ,คีย์นอก(Foreign key),คีย์คู่แข่ง(Candidate key)

คีย์หลักคือ แอททริบิวท์ที่แต่ละเรคคอร์ด(Record)ในตารางมีเพียงค่าเดียวและไม่เป็นค่าว่าง คีย์หลักสามารถประกอบไปด้วยแอททริบิวท์เดียวหรือหลายแอททริบิวท์รวมกัน ซึ่งในกรณีที่มีหลายแอททริบิวท์รวมกันเราจะเรียกว่า Composite primary key และเมื่อรู้คีย์หลัก ก็จะสามารถเข้าถึงแอททริบิวท์อื่นได้ อาจกล่าวได้ว่าคีย์หลักเป็นตัวแทนของเรคคอร์ด

คีย์นอกคือ แอททริบิวท์ที่แสดงความสัมพันธ์ระหว่างตารางหรือentity โดยคีย์นอกจะเป็นคีย์หลักของอีกตาราง คีย์นอกจึงเป็นตัวเชื่อมระหว่างตาราง

ความสัมพันธ์จะมีอยู่ 3 ลักษณะคือ one-to-one, one-to-many และ many-to-many (Roiger and Geatz 2003 : 181)ตัวอย่างเช่น สามี-ภรรยา จะเป็นความสัมพันธ์แบบ one-to-one, พ่อ-ลูก เป็นความสัมพันธ์แบบ one-to-many เนื่องจากพ่อหนึ่งคนสามารถมีลูกได้หลายคน และ ครู-ลูกศิษย์ เป็นความสัมพันธ์แบบ many-to-many เนื่องจากครู 1 คนสามารถมีลูกศิษย์ได้หลายคนและลูกศิษย์ 1 คนสามารถมีครูได้หลายคนเช่นกัน



ภาพที่ 8 ตัวอย่าง ER Diagram

การที่ ER Diagram จะสมบูรณ์นั้น จะต้องมีการวิเคราะห์ความสัมพันธ์ได้ดี หลักของการวิเคราะห์คือการทำนอร์แมลไลเซชั่น(Normalization)ทำให้ตัวแบบอยู่ในรูปแบบบรรทัดฐาน (Normal form) การทำนอร์แมลไลเซชั่นมีรูปแบบการออกแบบอยู่ 6 รูปแบบซึ่งจะเป็นการลดความซ้ำซ้อนของข้อมูล โดยการทำนอร์แมลไลเซชั่นนั้นไม่จำเป็นต้องทำให้ครบทั้ง 6 รูปแบบ จะขึ้นอยู่กับลักษณะของงานและผู้ทำการวิเคราะห์ โดยทั่วไปแล้วรูปแบบบรรทัดฐานที่นิยมใช้ในปัจจุบันมีอยู่ 3 รูปแบบและสามารถนิยามได้ดังนี้

1. รูปแบบบรรทัดฐานขั้นที่ 1 (First normal form หรือ 1NF) คือทุกแอททริบิวท์จะต้องมีค่าเดียว

2. รูปแบบบรรทัดฐานขั้นที่ 2 (Second normal form หรือ 2NF) ต้องเป็น 1NF และทุกแอททริบิวท์ที่ไม่ใช่คีย์หลักต้องขึ้นอยู่กับทุกคีย์หลักทั้งหมด หรืออีกนัยหนึ่งคือทุกแอททริบิวท์ที่ไม่ใช่คีย์ต้องไม่ขึ้นอยู่กับเพียงบางส่วนของคีย์หลัก

3. รูปแบบบรรทัดฐานขั้นที่ 3 (Third normal form หรือ 3NF) ต้องเป็น 2NF และ ไม่มีแอททริบิวท์ที่ไม่เป็นคีย์หลักขึ้นอยู่กับคีย์หลักอื่น หรืออีกนัยหนึ่งทุกแอททริบิวท์ที่ไม่ใช่คีย์หลักไม่มีคุณสมบัติในการกำหนดค่าของแอททริบิวท์อื่นที่ไม่ใช่คีย์หลัก

ตัวอย่างการทำนอร์แมลไลเซชั่น

จากข้อมูลการลงทะเบียนซึ่งมีแอททริบิวท์รหัสนักศึกษา ชื่อนักศึกษา รหัสอาจารย์ ชื่ออาจารย์ที่ปรึกษา รหัสวิชา ชื่อวิชาที่ลงทะเบียน หมู่เรียน หน่วยกิต แสดงดังตารางที่ 1

ตารางที่ 1 ข้อมูลตัวอย่างการลงทะเบียน

รหัสนักศึกษา	ชื่อนักศึกษา	รหัสอาจารย์	ชื่ออาจารย์	รหัสวิชา	ชื่อวิชาที่ลงทะเบียน	หมู่เรียน	หน่วยกิต
45308301	สมชาย พลจันทร์	K1059	สัมพันธ์ เย็นสำราญ	729101	เศรษฐศาสตร์เบื้องต้น	01	3
				729111	คณิตศาสตร์และสถิติ	01	3
				999211	คอมพิวเตอร์เบื้องต้น	01	3
45308302	สุทิศา พินิจไพทูรย์	K1011	ศิริภัทรา เหมือนมาลัย	729111	คณิตศาสตร์และสถิติ	02	3
				999211	คอมพิวเตอร์เบื้องต้น	02	3
				729104	การจัดการการเงิน	01	3

จากตารางข้อมูลการลงทะเบียนจะเห็นได้ว่านักเรียนหนึ่งคนสามารถลงทะเบียนได้มากกว่า 1 วิชาและโดยบางวิชาอาจเปิดสอนได้มากกว่า 1 หมู่เรียน จากนิยาม 1NF จะเห็นว่าตารางการลงทะเบียนยังไม่เป็น 1NF เนื่องจากมีบางแอททริบิวท์มีข้อมูลอยู่หลายค่า จะทำการนอร์แมลไลเซชั่น ให้เป็น 1NF จะแสดงได้ดังตารางที่ 2

ตารางที่ 2 ตัวอย่างการทำ 1NF

รหัสนักศึกษา	ชื่อนักศึกษา	รหัสอาจารย์	ชื่ออาจารย์	รหัสวิชา	ชื่อวิชาที่ลงทะเบียน	หมู่เรียน	หน่วยกิต
45308301	สมชาย พลจันทร์	K1059	สัมพันธ์ เย็นสำราญ	729101	เศรษฐศาสตร์เบื้องต้น	01	3
45308301	สมชาย พลจันทร์	K1059	สัมพันธ์ เย็นสำราญ	729111	คณิตศาสตร์และสถิติ	01	3
45308301	สมชาย พลจันทร์	K1059	สัมพันธ์ เย็นสำราญ	999211	คอมพิวเตอร์เบื้องต้น	01	3
45308302	สุทิศา พินิจไพทูรย์	K1011	ศิริภัทรา เหมือนมาลัย	729111	คณิตศาสตร์และสถิติ	02	3
45308302	สุทิศา พินิจไพทูรย์	K1011	ศิริภัทรา เหมือนมาลัย	999211	คอมพิวเตอร์เบื้องต้น	02	3
45308302	สุทิศา พินิจไพทูรย์	K1011	ศิริภัทรา เหมือนมาลัย	729104	การจัดการการเงิน	01	3

จาก 1NF ตรวจสอบพบว่าต้องใช้แอททริบิวท์รหัสนักศึกษาและรหัสวิชาเป็นคีย์หลัก จะเห็นได้ว่ามีแอททริบิวท์บางแอททริบิวท์ขึ้นอยู่กับส่วนหนึ่งของคีย์ เช่น ชื่อของนักศึกษาขึ้นอยู่กับรหัสนักศึกษา ซึ่งเป็นส่วนหนึ่งของคีย์หลัก ดังนั้นยังไม่เป็น 2NF จะทำการนอร์แมลไลเซชั่น ให้เป็น 2NF จะแสดงได้ดังตารางที่ 3 , 4 และ 5 ตามลำดับ

ตารางที่ 3 ตัวอย่างการทำ 2NF

รหัสนักศึกษา	ชื่อนักศึกษา	รหัสอาจารย์	ชื่ออาจารย์
45308301	สมชาย พลจันทร์	K1059	สัมพันธ์ เย็นสำราญ
45308302	สุทิศา พินิจไพทูรย์	K1011	ศิริภัทรา เหมือนมาลัย

ตารางที่ 4 ตัวอย่างการทำ 2NF

รหัสวิชา	ชื่อวิชาที่ลงทะเบียน	หน่วยกิต
729101	เศรษฐศาสตร์เบื้องต้น	3
729111	คณิตศาสตร์และสถิติ	3
999211	คอมพิวเตอร์เบื้องต้น	3
729104	การจัดการการเงิน	3

ตารางที่ 5 ตัวอย่างการทำ 2NF

รหัสนักศึกษา	รหัสวิชา	หมู่เรียน
45308301	729101	01
45308301	729111	01
45308301	999211	01
45308302	729111	02
45308302	999211	02
45308302	729104	01

จากตารางที่ 3 , 4 และ 5 ซึ่งเป็น 2NF จะพบว่าในตารางที่ 3 พบว่าจะใช้รหัสนักศึกษาเป็นคีย์หลัก ตารางที่ 4 จะใช้รหัสวิชาเป็นคีย์หลัก และ ตารางที่ 5 จะใช้รหัสนักศึกษาและรหัสวิชาเป็นคีย์หลัก จากการตรวจสอบในตารางที่ 3 พบชื่ออาจารย์ที่ปรึกษาขึ้นอยู่กับรหัสอาจารย์ ดังนั้นยังไม่เป็น 3NF จะทำการนอร์แมลไลเซชั่น ให้เป็น 3NF ดังแสดงในตารางที่ 6 และ 7

ตาราที่ 6 ตัวอย่างการทำ 3NF

รหัสนักศึกษา	ชื่อนักศึกษา	รหัสอาจารย์
45308301	สมชาย พลจันทร์	K1059
45308302	สุทิศา พินิจไพฑูรย์	K1011

ตารางที่ 7 ตัวอย่างการทำ 3NF

รหัสอาจารย์	ชื่ออาจารย์
K1059	สัมพันธ์ เย็นสำรา
K1011	ศิริภัทรา เหมือนมาลัย

จากการทำนอร์แมลไลเซชั่น เราสามารถเขียนความสัมพันธ์ในการลงทะเบียน รูปของฟังก์ชันได้ดังนี้

รหัสนักศึกษา -> ชื่อนักศึกษา , รหัสอาจารย์

รหัสอาจารย์ -> ชื่ออาจารย์

รหัสวิชา -> ชื่อวิชา , หน่วยกิต

รหัสนักศึกษา,รหัสวิชา -> หมู่เรียน

วัตถุประสงค์ในการทำนอร์แมลไลเซชั่นคือการลดความซ้ำซ้อนของข้อมูล และจากการกระทำดังกล่าวมีประโยชน์อื่นในเรื่องต่างๆด้วยเช่น ทำให้ประหยัดเนื้อที่ในการจัดเก็บ ทำให้ลดปัญหาข้อมูลขาดความถูกต้อง(Data integrity) ลดปัญหาที่เกิดที่จากการปรับปรุง เพิ่มเติม และลบข้อมูล

2.1.2 ตัวแบบเชิงมิติ(Dimensional Modeling)

ตัวแบบเชิงมิติเป็นเทคนิคที่ใช้สำหรับช่วยในการวิเคราะห์ข้อมูลเป็นหลัก โดยตัวแบบเชิงมิติมุ่งประเด็นไปที่ข้อมูลที่เป็นตัวเลขเช่น ยอดขาย จำนวนที่ขาย น้ำหนัก เป็นต้น การใช้ตัวแบบเชิงมิติช่วยในการสร้างคลังข้อมูลจะมีความง่ายต่อการสร้างคลังข้อมูลและเข้าใจง่ายกว่าการใช้ตัวแบบเชิงสัมพันธ์ แต่ตัวแบบเชิงมิตินั้นก็ต้องอาศัยความเข้าใจในเรื่องความสัมพันธ์ของตัวแบบเชิงสัมพันธ์เช่นเดียวกัน ตัวแบบที่จะใช้มีหลักพื้นฐานอยู่ 3 ชนิดคือ ตัววัด(Measures) ข้อเท็จจริง(Facts) และมิติ(Dimensions) ซึ่งจะประกอบกันเป็นลักษณะของ Cube ด้วยลักษณะนี้จึงช่วยให้แสดงข้อมูลที่อยู่ในฐานข้อมูลตามความต้องการผู้ใช้ได้ดี

ข้อเท็จจริง(Facts) เป็นชุดของความสัมพันธ์ของรายการข้อมูล ตัววัด และเนื้อหาข้อมูล ที่แสดงถึงข้อเท็จจริงที่เกิดขึ้นในรายการธุรกิจ ธุรกรรม หรือเหตุการณ์ที่สามารถช่วยในการวิเคราะห์ข้อมูลทางธุรกิจหรือการดำเนินการทางธุรกิจ เช่น เมื่อต้องการดูการขายของบริษัทแบ่งตามคลังสินค้าแบ่งตามไตรมาส ซึ่งจะทำให้สามารถทราบถึงรายได้หรือกำไรสุทธิได้ นั้นคือข้อเท็จจริงที่เราสนใจจากข้อมูลที่ต้องการแสดงให้เห็น

มิติ(Dimension) เป็นมุมมองของตัววัด เป็นส่วนช่วยในการวิเคราะห์ข้อเท็จจริงที่เราสนใจ เนื่องจากมิติช่วยให้เราเห็นถึงความแตกต่างของข้อเท็จจริงนั้นๆ ตัวอย่างเช่น ยอดขายรวมสามารถแสดงมุมมองของคลัง เมือง หรือพื้นที่ในการขาย ซึ่งทั้งสามระดับนั้นก็เป็นสมาชิกของมิติภูมิศาสตร์ที่ใช้ในการขาย ในตัวแบบเชิงมิตินั้นข้อมูลทุกข้อมูลที่มีค่าเดียวกันในมิติเดียวกันนั้นจะถูกรวมกันหรือถูกสรุปข้อมูลให้เหลือเป็นเพียงค่าเดียวในแต่ละมิติ

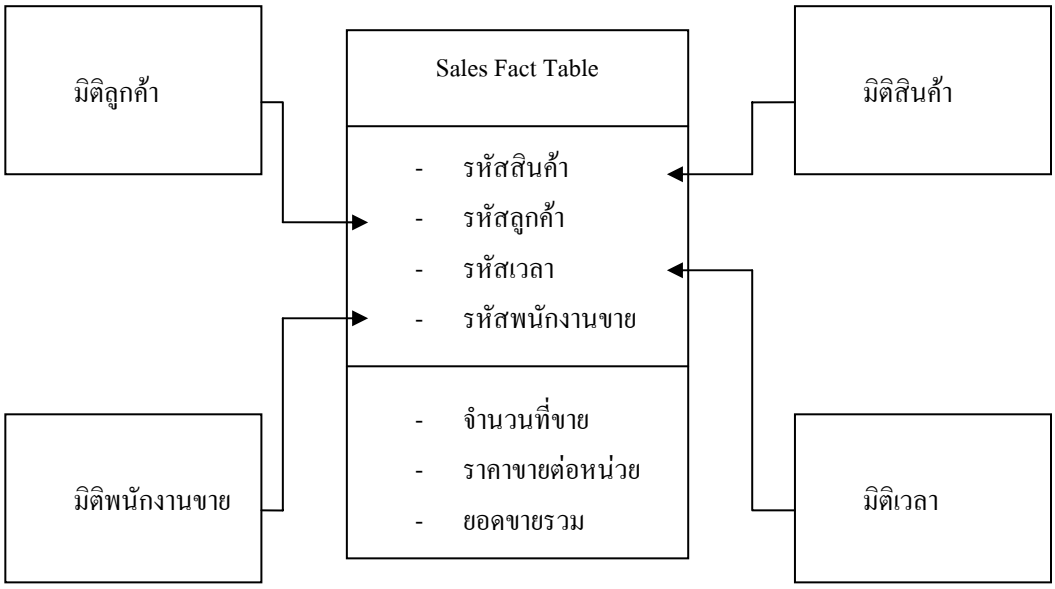
มิติ(Dimension) ตัววัด(Measure)

	ไตรมาส	ชนิดสินค้า	จังหวัด	ยอดขาย(B)
ข้อเท็จจริง ที่ 1 →	ไตรมาส 1 ,2548	เสื้อเด็ก	กรุงเทพ	560000
ข้อเท็จจริง ที่ 2 →	ไตรมาส 2 ,2548	กระโปรง	นครปฐม	450000
ข้อเท็จจริง ที่ 3 →	ไตรมาส 3, 2548	กางเกง	เชียงใหม่	790000

ภาพที่ 9 ความสัมพันธ์ระหว่าง ข้อเท็จจริง มิติ และตัววัด

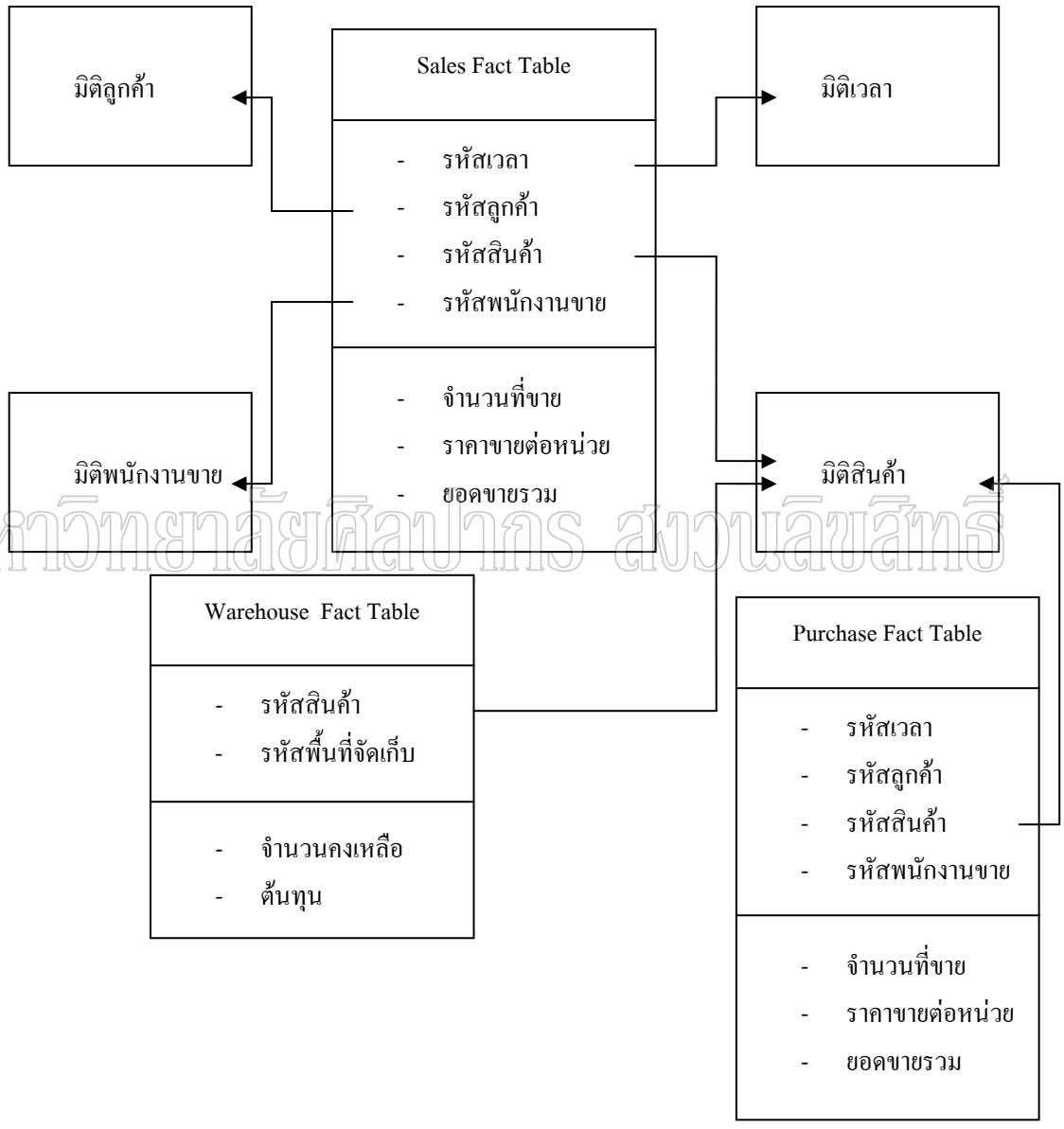
จากรูปเป็นการแสดงให้เห็นถึงข้อมูลการขาย ที่อาศัยการออกแบบโดยใช้ตัวแบบเชิงมิติ โดยมี ยอดขายเป็นตัววัด โดยมี ไตรมาส ชนิดสินค้า จังหวัดเป็นมิติ ประกอบกันเป็นข้อเท็จจริงที่แสดงตามดังรูป ซึ่งเรียกตารางในภาพที่9 ว่า ตารางข้อเท็จจริง(Fact table)โดย Schema ที่ใช้มีอยู่ในคลังข้อมูลมีอยู่ 3 แบบด้วยกันคือ

1. Star schema เป็น Schema ที่เข้าใจง่ายและเป็นที่นิยมโดยจะมี Fact table อยู่ตรงกลางและมีตารางที่เก็บคำอธิบายมิติซึ่งเรียกว่า Dimension table มีความสัมพันธ์อยู่รายรอบ ซึ่งจะมี Fact table เท่านั้นที่ใช้ Multiple join ไปยัง Dimension table ที่อยู่รายรอบ โดยการเชื่อมของ Dimension table อื่นๆไปยัง Fact table จะมีลักษณะเป็น Single join (สุนีย์ พงษ์พินิจภิญโญ ม.ป.ป. : 423) เมื่อมองลักษณะของDiagram แล้วจะคล้ายกับดาว



ภาพที่ 10 Star schema

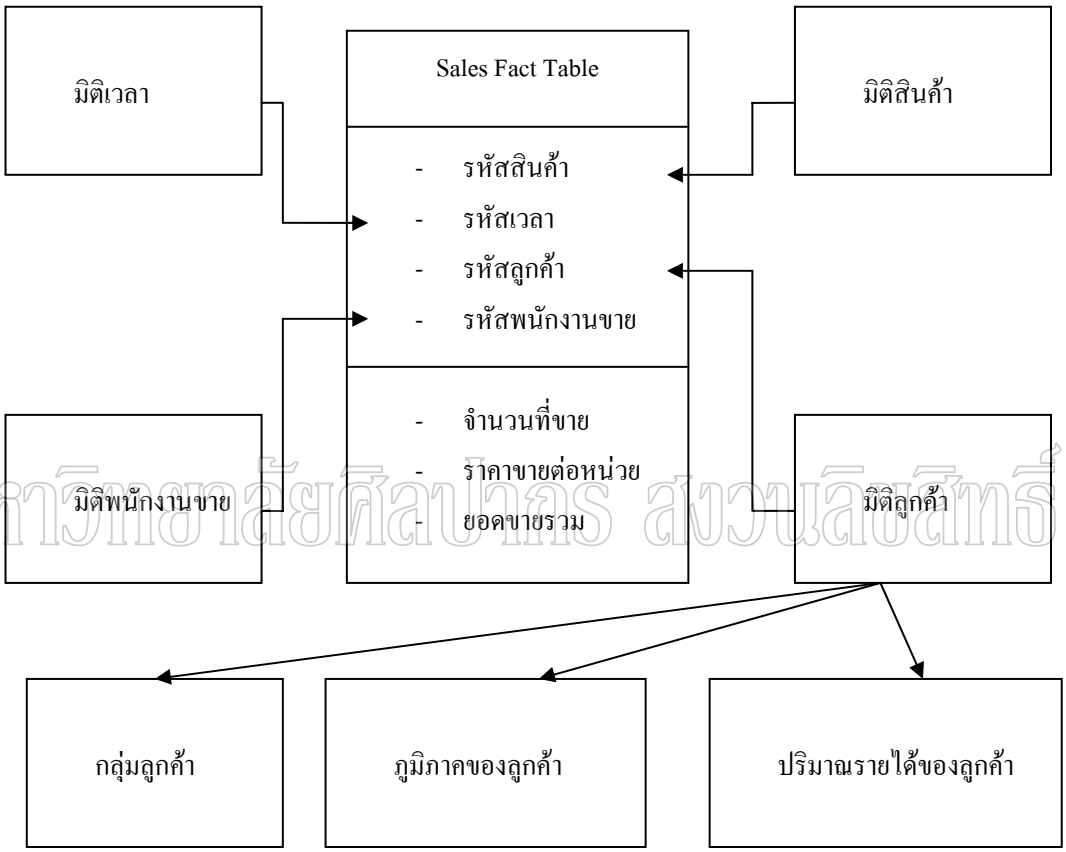
2 . Constellation schema จะมีลักษณะเหมือนกับ Star schema แตกต่างตรงที่จะมี Fact table มากกว่าหนึ่ง ดังนั้น Dimension table จะสามารถเชื่อมกับ Fact table ในลักษณะ Multiple join แต่ Constellation schema ไม่เป็นที่นิยมเนื่องจากมีความซับซ้อนมากทำให้การเรียกใช้ข้อมูลค่อนข้างยุ่งยาก (Youness 2005 : 68)



ภาพที่ 11 Constellation schema

3. Snowflake schema จากที่ทราบแล้วว่า Star schema เป็น Schema ที่เข้าใจง่ายที่จะใช้แสดงข้อมูลในการวิเคราะห์ อย่างไรก็ตามบางครั้งเมื่อข้อมูลมากขึ้นทำให้ข้อมูลใน Dimension table

มากขึ้น เพื่อเป็นการทำให้การวิเคราะห์ข้อมูลมีประสิทธิภาพมากขึ้น จึงทำการแบ่งข้อมูลของ Dimension table ออกมา (Youness 2005 : 70)จึงเป็นลักษณะของ Snowflake Schema ดังนั้น กล่าวได้ว่า Snowflake schema จะแตกต่างกับ Star schema ตรงที่ Dimension table จะถูกแบ่งย่อยออกมาและมีการเชื่อมความสัมพันธ์กับ Dimension table เดิม



ภาพที่ 12 Snowflake schema

เนื่องจากคลังข้อมูลจะเป็นการเก็บรวบรวมข้อมูลธุรกรรมจากหน่วยงานต่าง ๆ เพื่อช่วยในการวิเคราะห์จึงจำเป็นต้องมีส่วนใดส่วนหนึ่งในคลังข้อมูลสำหรับใช้พรรณนาลักษณะของข้อมูล แหล่งที่มา รูปแบบ ข้อกำหนดในการใช้งาน รวมทั้งการควบคุมข้อมูลและกระบวนการต่างๆที่มีอยู่ในคลังข้อมูลซึ่งส่วนนี้จะเรียกว่าเม็ตตาเดตา(Metadata) (กิตติพงศ์ กลมกล่อม 2548 : 168-169) ตัวอย่างข้อมูลในเม็ตตาเดตาคือ

- คำอธิบายและแหล่งที่มาของแต่ละฟิลด์

- ควรจะโหลดหรือปรับปรุงข้อมูลเมื่อใด
- ความปลอดภัยและสิทธิที่เข้าถึงข้อมูล
- การเชื่อมโยงกับระบบอื่น เป็นต้น

2.2 การโอนถ่ายข้อมูล

การโอนถ่ายข้อมูลมีความสำคัญต่อคลังข้อมูลเนื่องจากคลังข้อมูลจะทำการรวบรวมข้อมูลจากแหล่งข้อมูล(Data source)ทั้งภายในและภายนอกองค์กร ดังนั้นจึงต้องมีวิธีการในการรวบรวมข้อมูลทั้งหลายเหล่านั้น ซึ่งจะมีการโอนถ่ายข้อมูลจากแหล่งข้อมูลต่างๆสู่คลังข้อมูล และการโอนถ่ายข้อมูลจากคลังข้อมูลสู่ Data marts การโอนถ่ายข้อมูลจะมี 2 ส่วนคือ

1. Data acquisition เป็นส่วนที่รับและเตรียมข้อมูลก่อนที่จะนำข้อมูลเข้าคลังข้อมูลไม่ว่าจะเป็นข้อมูลมาจากแหล่งข้อมูลใด (กิตติพงศ์ กลมกล่อม 2548 : 118) หน้าที่หลักของ Data acquisition คือ

- ตรวจสอบข้อมูลเบื้องต้นพร้อมทั้งแก้ไขข้อมูลที่ผิดพลาดให้มีความถูกต้อง เช่น ข้อมูลที่เป็นตัวเลขต้องมีค่าเป็นตัวเลข
- ทำการกรองข้อมูลที่ไม่ต้องการหรือมีความผิดพลาดที่อาจจะทำให้คลังข้อมูลเกิดความผิดพลาดในขณะทำงาน เช่น ไวรัสคอมพิวเตอร์
- การรักษาความปลอดภัยในการส่งข้อมูล เช่น การตรวจสอบสิทธิผู้ส่งข้อมูล
- แจ้งเตือนข้อมูลที่ผิดพลาด

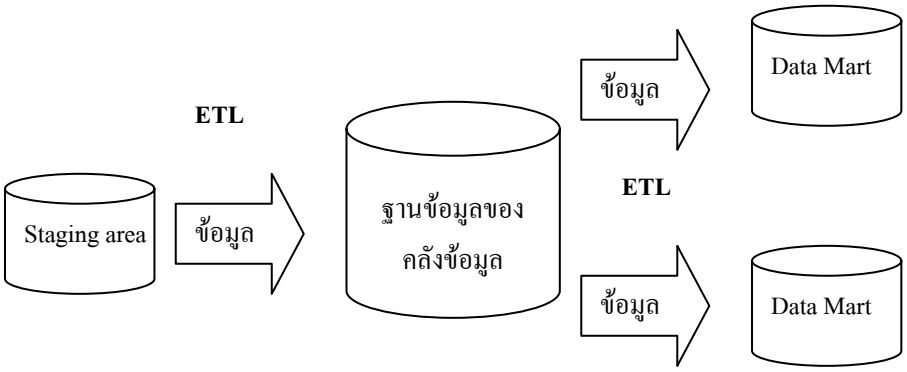
2. Data staging area เป็นส่วนแปลงข้อมูลที่ได้รับจาก Data acquisition ให้อยู่ในรูปแบบที่คลังข้อมูลกำหนด หรือตามตัวแบบคลังข้อมูล อีกทั้งยังมีกระบวนต่างๆ ที่อยู่ในส่วนของ Data staging area(กิตติพงศ์ กลมกล่อม 2548 : 144-145) เช่น การเป็นที่พักข้อมูลและสำรองข้อมูลชั่วคราว (Temporary backup) ในระหว่างการโอนถ่ายข้อมูลก่อนที่จะนำข้อมูลเข้าสู่คลังข้อมูล กระบวนการทำความสะอาดข้อมูล (Data cleansing) และการตรวจสอบข้อมูล ซึ่งการตรวจสอบข้อมูลนั้นจะแตกต่างกับส่วนของ Data acquisition เนื่องจาก Data staging area สามารถติดต่อได้ทั้งในส่วนของ Data acquisition และคลังข้อมูล แต่ Data acquisition ไม่สามารถติดต่อกับส่วนของคลังข้อมูล ดังนั้นการตรวจสอบข้อมูลในส่วนของ Data staging area เป็นการตรวจสอบความถูกต้องและยังสามารถตรวจสอบความสอดคล้องของข้อมูลที่มีอยู่ในคลังข้อมูลเดิม เช่น การตรวจสอบรหัสสินค้าที่สอดคล้องหรือมีอยู่จริงในคลังข้อมูล

กระบวนการที่จะนำข้อมูลจากส่วนหนึ่งไปยังอีกส่วนหนึ่งเรียกว่า ETL (Extraction ,Transformation and Load) ซึ่งประกอบไปด้วย 3 กระบวนการ คือ

1. Extract คือกระบวนการของการดึงข้อมูลออกจากแหล่งข้อมูล

2. Transform คือ กระบวนการแปลงข้อมูลจากโครงสร้างเดิมจากแหล่งข้อมูล(Source) ให้อยู่ในรูปแบบโครงสร้างข้อมูลปลายทาง(Destination)

3. Load คือ กระบวนการนำของมูลที่แปลงรูปแบบแล้วเข้าสู่แหล่งข้อมูลปลายทาง



ภาพที่ 13 ETL

โดยภาษาที่ใช้ในการจัดการข้อมูลในฐานข้อมูลนั้นอาจจะใช้ ภาษา SQL (Structured Query Language) ช่วยในการจัดการข้อมูล โดยภาษา SQL เป็นภาษาทางด้านฐานข้อมูล แบ่งออกเป็น 3 กลุ่ม คือ

1. ภาษาที่ใช้สำหรับนิยามข้อมูล (Data Definition Language :DDL) จะกลุ่มของคำสั่งที่ใช้ในการกำหนดโครงสร้างต่าง ๆ ของฐานข้อมูล
2. ภาษาที่ใช้ในการจัดการข้อมูล(Data Manipulation Language :DML) จะเป็นกลุ่มคำสั่งที่ใช้ในการเรียกดู เปลี่ยนแปลง เพิ่มหรือลบข้อมูล
3. ภาษาที่ใช้ในการควบคุม(Data Control Language :DCL) เป็นกลุ่มคำสั่งที่ใช้ในการควบคุมข้อมูล เช่น การเรียกใช้ข้อมูลพร้อมกัน การกำหนดความปลอดภัยของข้อมูล

2.3 การสร้างผลลัพธ์หรือการวิเคราะห์ข้อมูลในคลังข้อมูล

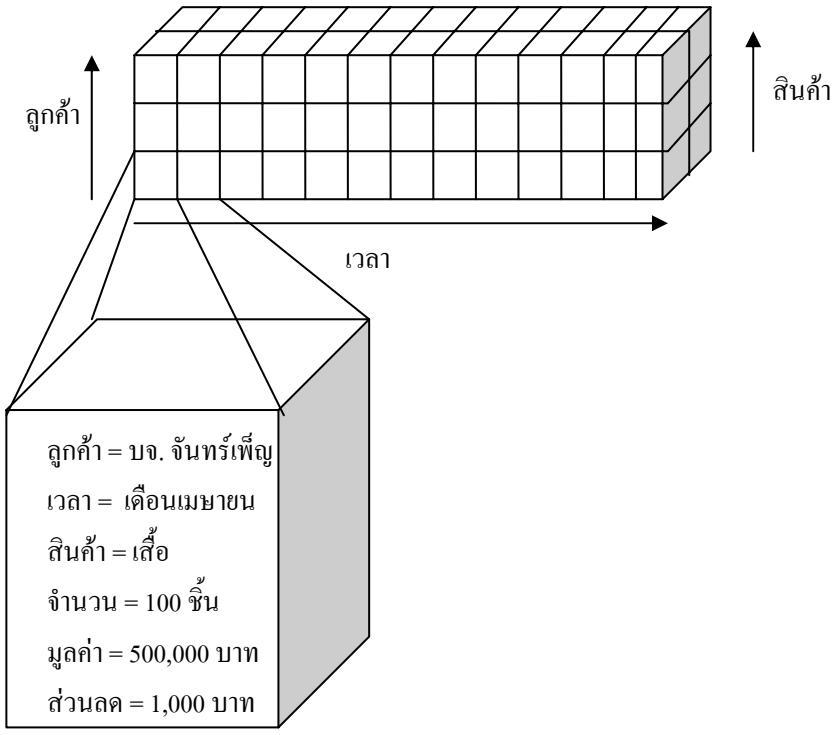
โดยทั่วไปแล้วจะมีการดำเนินการที่นิยมทำอยู่ 3 รูปแบบด้วยกัน คือ รายงานและสอบถาม (Report and Query) , วิเคราะห์แบบหลายมิติ(Multidimensional data analysis) หรือการทำการประมวลผลในเชิงวิเคราะห์แบบออนไลน์(On-Line-Analytic Processing หรือ OLAP),และเหมืองข้อมูล(Data mining) โดยเหมืองข้อมูลจะกล่าวในหัวข้อถัดไป

ข้อแตกต่างระหว่างรายงานและสอบถามกับ OLAP คือรายงานและสอบถามจะทำการประมวลผลจากฐานข้อมูลทุกครั้งที่การเรียกใช้แต่ OLAP จะทำการประมวลผลไว้ก่อนในช่วงเวลา

ที่ข้อมูลไม่ได้ถูกใช้งานหรือข้อมูลที่มีการสรุปไว้แล้วในCubeเมื่อเรียกดูข้อมูลจะมีประสิทธิภาพและรวดเร็วกว่า ในบางครั้งเครื่องมือที่ใช้สร้างระบบ On-Line Analytic Processing(OLAP) จะนำเอาวิธีการเหมืองข้อมูลเข้าไปรวมอยู่ด้วยสำหรับเป็นเครื่องมือชนิดหนึ่งในการวิเคราะห์

รายงานและสอบถาม จะเป็นการใช้โปรแกรมหรือระบบที่เรียกว่า “ระบบสร้างรายงาน (Report generator)”เพื่อใช้ข้อมูลที่เกิดจากการปฏิบัติงานหรือในคลังข้อมูลมาประมวลผล เพื่อประโยชน์ในการตัดสินใจ(กิตติพงศ์ กลมกล่อม 2548 :11) โดยรายงานที่ได้ไม่มีความซับซ้อนมากนัก กล่าวคือเป็นวิธีการที่ง่ายที่สุดในการวิเคราะห์ข้อมูลในคลังข้อมูล

วิเคราะห์แบบหลายมิติ(Multidimensional data analysis) หรือการทำการประมวลผลในเชิงวิเคราะห์แบบออนไลน์(On-Line-Analytic Processing หรือ OLAP)ซึ่งต่อไปเราจะเรียกว่า OLAP ซึ่งเป็นวิธีการสอบถามที่สนับสนุนการวิเคราะห์ข้อมูลในสภาพแวดล้อมหลายมิติ ที่ประกอบไปด้วยมิติ(Dimension)และตารางข้อเท็จจริง(Fact Table) ซึ่งตารางข้อเท็จจริง คือข้อเท็จจริงที่เราต้องการวิเคราะห์เช่นยอดขาย ผลกำไร จำนวนสินค้าที่มีอยู่ในปัจจุบัน เป็นต้น ในตารางข้อเท็จจริง จะประกอบไปด้วย Measure เป็นข้อมูลที่ต้องการวัดทั้งในเชิงปริมาณ (Quantitative)และเชิงคุณภาพ(Qualitative) กับข้อมูลที่ใช้เชื่อมความสัมพันธ์กับมิติ (Dimension)โดยที่ข้อมูลจะถูกเก็บอยู่ในรูปของลูกบาศก์ที่มีหลายมิติซึ่งเรียกว่า Cube



ภาพที่ 14 ลักษณะของ Cube

การใช้ระบบ OLAP ช่วยตอบสนองความต้องการของผู้ใช้ได้รวดเร็วขึ้น สามารถใช้ได้กับปัญหาทางธุรกิจที่เกิดขึ้นจริง ช่วยในการตัดสินใจของผู้บริหารในองค์กรและยังใช้ทรัพยากรมนุษย์ได้อย่างมีประสิทธิภาพ (Youness 2005 :10) จึงทำให้เพิ่มรายได้และผลกำไรมากขึ้น ตัวอย่างเช่น แผนกขายใช้ OLAP เป็นเครื่องในการทำการวิเคราะห์ขาย ช่วยให้แผนกขายสามารถขายโดยใช้เทคนิคการขายที่ดีสุด และทราบว่าสินค้าใดที่จะขายได้มากกว่าสินค้าชนิดอื่น หรืออย่างเช่นแผนกการตลาดใช้ OLAP เป็นเครื่องมือในการวิเคราะห์การส่งเสริมการขาย วิเคราะห์ลูกค้า เป็นต้น โดยลักษณะเด่นของ OLAP มีลักษณะดังนี้

1. สามารถมองข้อมูลได้หลายมิติ(Multidimensional views of data หรือ Data cubes) โดยปกติแล้วตัวแบบทางธุรกิจสามารถมองข้อมูลได้หลายมิติ(Youness 2005 :11) ตัวอย่างเช่น การขายเราสามารถมองยอดขายในมิติของเวลา สินค้า สถานที่ ลูกค้า เป็นต้น โดยที่มิติของเวลาเราอาจจะดูข้อมูลเป็นรายเดือน รายไตรมาส รายปี ในมิติสถานที่สามารถดูข้อมูลตามลักษณะภูมิศาสตร์ มิติของสินค้าสามารถดูข้อมูลตามกลุ่มของสินค้า ในการมองข้อมูลหลายมิตินั้นจะอ้างถึง data cube ซึ่ง cube ในตัวแบบทางธุรกิจอาจจะมิติได้มากกว่าสามมิติ

2. สามารถที่จะคำนวณข้อมูลได้มากขึ้น(Calculation-intensive) เนื่องจาก OLAP จะทำการสรุปข้อมูลตามลำดับชั้น ทำให้สามารถทำการคำนวณที่ซับซ้อนได้มากขึ้น(Youness 2005 :12) เช่นเปอร์เซ็นต์ของจำนวนรวมทั้งหมด ค่าเฉลี่ยแบบเคลื่อนที่ เปอร์เซ็นต์การเจริญเติบโต ซึ่งOLAP ถูกออกแบบมาให้รองรับการคำนวณที่ซับซ้อน ซึ่งจะเป็นผลประโยชน์ต่อการตัดสินใจมากขึ้น โดยปกติแล้วระบบ OLTP(On-Line Transaction Processing)เป็นระบบที่ใช้ในการรวบรวมและจัดการข้อมูล แต่ระบบOLAP จะเป็นระบบที่ใช้สร้างสารสนเทศจากข้อมูลที่ถูกรวบรวมซึ่งจะทำให้เกิดองค์ความรู้ใหม่

3. สามารถคำนวณข้อมูลตามเวลาได้ (Time-Intelligence) กล่าวคือ เวลาเป็นมิติที่นิยมใช้มากในระบบ OLAP มิติของเวลานั้นสามารถใช้ในการเปรียบเทียบประสิทธิภาพของการดำเนินการทางธุรกิจ (Youness 2005 :12) ตัวอย่างเช่น ในการพิจารณาการปฏิบัติงานของเดือนปัจจุบันกับการปฏิบัติงานของเดือนที่ผ่านมา หรือ การเปรียบเทียบผลกำไรขององค์กรในไตรมาสล่าสุดกับไตรมาสเดียวกันในปีที่ผ่านมา เป็นต้น

2.4 การดำเนินการกับ OLAP

1. Roll up และ Drill down

เนื่องจาก OLAP จะมีการสรุปข้อมูลตามมิติต่างๆ ที่เราสนใจและข้อมูลตามมิตินั้นจะมีการเก็บอยู่ในลักษณะของลำดับชั้น ดังนั้น Roll up และ Drill down คือการเปลี่ยนแปลงระดับความละเอียด

ของลำดับชั้นของมิติในการพิจารณาข้อมูล โดยการดำเนินการนี้จะใช้กับ Snowflakes schema เป็นส่วนใหญ่ (กิตติตพงศ์ กลมกล่อม 2548 : 110)

โดยการ Roll up จะเป็นการเปลี่ยนแปลงระดับความละเอียดจากระดับที่เล็กไปสู่ระดับที่ใหญ่กว่า และ Drill down จะเป็นการเปลี่ยนแปลงในตรงกันข้ามกัน กล่าวคือจะเป็นการเปลี่ยนแปลงระดับความละเอียดจากระดับที่ใหญ่ไปสู่ระดับที่เล็กกว่า ตัวอย่างการทำ Roll up และ Drill down พิจารณาตารางข้อเท็จจริงของการขายแยกตามภูมิภาค (Region) ร้านค้า(Shop) ประเภทสินค้า (Product type) และสินค้า (Product) ซึ่งแสดงดังตารางที่ 8

ตารางที่ 8 ตัวอย่างตารางข้อเท็จจริงของการขาย(Sale Fact Table)

Region	Shop	Product type	Product	Amount
ภาคเหนือ	A	เสื้อ	เสื้อแขนยาว	100,000
ภาคเหนือ	A	เสื้อ	เสื้อแขนยาว	200,000
ภาคใต้	B	กางเกง	กางเกงขายาว	200,000
ภาคใต้	B	กางเกง	กางเกงขาสั้น	250,000
ภาคใต้	B	กระโปรง	กระโปรงสั้น	150,000
ภาคใต้	C	กระโปรง	กระโปรงสั้น	300,000

จาก Sale fact table สามารถสร้าง Cube ได้ดังนี้

ตารางที่ 9 ตัวอย่าง Cube ที่ได้จากตารางข้อเท็จจริงของการขาย

	Product Type	กระโปรง	กางเกง	กางเกง	เสื้อ
	Product	กระโปรงสั้น	กางเกงขายาว	กางเกงขาสั้น	เสื้อแขนยาว
Region	Shop				
ภาคใต้	B	150,000	200,000	250,000	
ภาคใต้	C	300,000			
ภาคเหนือ	A				300,000

จาก Cube ดังกล่าวจะแสดงการทำ Roll up และ Drill down ได้ดังนี้

Region	Sum
ภาคใต้	900,000
ภาคเหนือ	300,000

Drill down ⇩ ⇧ Roll up

	Product Type	กระโปรง	กางเกง	เสื้อ
Region				
ภาคใต้		450,000	450,000	
ภาคเหนือ				300,000

Drill down ⇩ ⇧ Roll up

	Product Type	กระโปรง	กางเกง	เสื้อ
Region	Shop			
ภาคใต้	B	150,000	450,000	
ภาคใต้	C	300,000		
ภาคเหนือ	A			300,000

Drill down ⇩ ⇧ Roll up

	Product Type	กระโปรง	กางเกง	กางเกง	เสื้อ
	Product	กระโปรงสั้น	กางเกงขายาว	กางเกงขาสั้น	เสื้อแขนยาว
Region	Shop				
ภาคใต้	B	150,000	200,000	250,000	
ภาคใต้	C	300,000			
ภาคเหนือ	A				300,000

ภาพที่ 15 การทำ Drill down และ Roll up

2. Slice

เป็นพิจารณาเฉพาะบางส่วนของ Cube ที่เราสนใจ เพราะในบางครั้งถ้า Cube มีขนาดใหญ่มากการพิจารณาข้อมูลทั้งหมดอาจจะทำให้ยากต่อการพิจารณา การวิเคราะห์โดยการแบ่ง Cube ออกมาเฉพาะบางส่วน จะช่วยให้ง่ายขึ้นต่อการพิจารณา จากตารางข้อเท็จจริงของการขาย ในตารางที่ 9 จะทำการ Slice เฉพาะข้อมูลที่สนใจดังนี้

ตารางที่ 10 ตัวอย่างการทำ Slice

	Product Type	กระโปรง	กางเกง	กางเกง
	Product	กระโปรงสั้น	กางเกงขายาว	กางเกงขาสั้น
Region	Shop			
ภาคใต้	B	150,000	200,000	250,000
ภาคใต้	C	300,000		

3. Dice

เป็นการพลิกแกนของมิติใน Cube ซึ่งเป็นตอบสนองมุมมองตามความต้องการของผู้วิเคราะห์ข้อมูล แสดงดังตัวอย่าง โดยใช้ Cube จากตารางข้อเท็จจริงของการขาย

ตารางที่ 11 ตัวอย่างการทำ Dice

Region	ภาคใต้	ภาคเหนือ
Shop		
B	600,000	
C	300,000	
A		300,000

3. การทำเหมืองข้อมูล(Data mining)

3.1 นิยามของเหมืองข้อมูล

เหมืองข้อมูลคือกระบวนการในการค้นหาความสัมพันธ์ รูปแบบ และแนวโน้มของข้อมูลจากข้อมูลที่ถูกเก็บรวบรวมไว้เป็นจำนวนมากโดยจะใช้เทคนิคทางสถิติหรือเทคนิคทางคณิตศาสตร์ (Berry and Linoff 2004 :7) นอกจากนี้แล้วเหมืองข้อมูลยังมีผู้ให้นิยามอีกหลายนิยาม เช่น เหมืองข้อมูล คือกลั่นกรองสารสนเทศที่อยู่ฐานข้อมูลขนาดใหญ่โดยใช้เทคนิคของ machine learning ,วิธีทางสถิติ ,วิธีทางคณิตศาสตร์ ,วิธีทางฐานข้อมูล และการแสดงข้อมูลในรูปแบบรายงานต่าง ๆ (Cabena and other : quoted in Daniel T. Larose 2003 : 2) หรือ เหมืองข้อมูลเป็นการค้นหาวิเคราะห์ หรือสร้างองค์ความรู้ใหม่จากข้อมูลขนาดใหญ่ซึ่งอาจจะเป็นการค้นหารูปแบบหรือกฎ โดยการใช้เทคนิคทางสถิติ ทางคณิตศาสตร์หรือเทคนิคทางวิทยาการคอมพิวเตอร์

จากนิยามการทำเหมืองข้อมูลจะช่วยให้สามารถได้สารสนเทศที่ซ่อนอยู่ในข้อมูล และยังเป็นการใช้ข้อมูลอย่างคุ้มค่า แต่การทำเหมืองข้อมูลเป็นวิธีการที่ค้อนข้างยุ่งยากจึงต้องนำโปรแกรมคอมพิวเตอร์เข้ามาช่วยในการวิเคราะห์เช่น SPSS , SAS ,Minitab ,Oracle ,Microsoft SQL Server หรือโปรแกรมทางสถิติและทางคณิตศาสตร์ต่างๆ ซึ่งในปัจจุบันโปรแกรมสำเร็จรูปมีการนำคุณสมบัติการทำเหมืองข้อมูลเพิ่มเติมเข้ามา ทั้งนี้ผู้ทำเหมืองข้อมูลยังสามารถพัฒนาโปรแกรมเองเพื่อเข้ามาช่วยวิเคราะห์เองได้ด้วย

3.2 วิวัฒนาการของเหมืองข้อมูล

เทคนิคเหมืองข้อมูลเป็นผลจากการวิจัยและกระบวนการพัฒนาผลผลิตเป็นเวลานาน โดยเริ่มระยะแรกในปี ค.ศ. 1960 เริ่มมีการนำคอมพิวเตอร์เข้ามาช่วยในการดำเนินการทางธุรกิจ ในการวิเคราะห์ข้อมูลระยะแรกเป็นของเทคนิค Data collection เป็นการใช้ข้อมูลที่เก็บรวบรวมมาวิเคราะห์ซึ่งเป็นการคำนวณข้อมูลอย่างง่ายเช่น การหาค่าเฉลี่ย หรือผลรวม สารสนเทศที่สร้างขึ้นมาในระยะนี้เป็นการตอบคำถามทางธุรกิจเช่น รายได้รวม หรือรายได้เฉลี่ยในแต่ระยะคาบเวลา ต่อมาในปี ค.ศ. 1980 เป็นระยะเป็นของเทคนิค Data access มีการใช้เทคโนโลยีฐานข้อมูลมาช่วยในการเก็บข้อมูล มีเริ่มมีการใช้ระบบการจัดการรายงานเป็นเครื่องมือในการช่วยสร้างสารสนเทศ ปี ค.ศ. 1990 เป็นระยะของเทคนิค Data navigation โดยเริ่มมีการใช้ ฐานข้อมูลเชิงหลายมิติ ,OLAP และคลังข้อมูลเพื่อเข้ามาตอบคำถามและปัญหาต่างๆ ที่ซับซ้อนขึ้น พร้อมทั้งให้ทันต่อเหตุการณ์ที่เกิดขึ้น เช่น ยอดขายรวมในแต่ละภูมิภาคแยกตามคาบระยะเวลาที่ต้องการ สุดท้ายในปี ค.ศ. 2000 เป็นระยะเริ่มใช้เทคนิค เหมืองข้อมูล(Data mining) ซึ่งเป็นวิเคราะห์ที่หาสารสนเทศที่ซ่อนอยู่ในข้อมูลโดยอาศัยข้อมูลจำนวนมาก

ตารางที่ 12 วิวัฒนาการของเหมืองข้อมูล

เทคนิค(ปีที่เริ่ม)	คำถามหรือปัญหาทางธุรกิจ	เทคโนโลยีที่เกิดขึ้น
Data collection(1960s)	ค่าเฉลี่ยของรายได้ห้าเดือนที่ผ่านมา	คอมพิวเตอร์,เทป,disks
Data Access (1980s)	ยอดขายที่นิวอิงแลนด์เมื่อเดือนเมษายน	ฐานข้อมูลเชิงสัมพันธ์,SQL,ODBC
Data Navigation(1990s)	ยอดขายที่นิวอิงแลนด์เมื่อเดือนเมษายนโดยแยกตามเมืองต่างๆ	OLAP,ฐานข้อมูลเชิงมิติ,คลังข้อมูล
เหมืองข้อมูล(2000)	อะไรน่าจะเกิดขึ้นจากการขายในบอสตันของเดือนหน้า เพราะอะไร	อัลกอริธึมชั้นสูง,ฐานข้อมูลขนาดใหญ่, การประมวลคอมพิวเตอร์หลายขั้นตอน

ที่มา : Chris Rygielski, Jyun-Cheng Wang and David C. Yen , “Data mining Techniques for Customer Relationship Management” Technology In Society 24 (2002):484.

3.3 ตัวแบบการดำเนินการของเหมืองข้อมูล

จากที่กล่าวมาแล้วว่าการทำเหมืองข้อมูลเป็นขั้นตอนที่สำคัญของ Knowledge Discovery in Data (KDD) ดังนั้นเราสามารถใช้ตัวแบบในการทำ KDDเข้ามาช่วย ซึ่ง KDD คือ กระบวนการในการกำหนดและการแสวงหารูปแบบที่ชัดเจน เป็นองค์ความรู้ใหม่ที่มีประโยชน์และเข้าใจได้จากสิ่งที่ซ่อนอยู่ในข้อมูล จึงมีลักษณะของการดำเนินการเหมือนกัน โดยตัวแบบการดำเนินการมีอยู่หลายตัวแบบด้วยกันแต่ที่นำเสนอนี้จะขอยกมาเพียง 3 ตัวแบบ คือ

3.3.1 ตัวแบบ nine step of KDD มี 9 ขั้นตอนดังนี้

1. กำหนดปัญหาและขอบเขตพร้อมทั้งศึกษาปัญหาที่เกิดขึ้น

2. สร้างเซตข้อมูลเป้าหมายที่จะใช้ในการ KDD โดยข้อมูลที่ใช้จะต้องใช้แก้ปัญหาหรือตอบคำถามที่กำหนดได้

3. Data Cleaning and preprocessing เป็นกระบวนการที่ใช้ในเตรียมและตรวจสอบข้อมูลให้ถูกต้อง ก่อนที่จะนำไปใช้งาน

4. Data reduction and projection เป็นขั้นตอนของการสรุปข้อมูลและลดความซ้ำซ้อนของข้อมูล

5. เลือกฟังก์ชันของเหมืองข้อมูลเพื่อใช้ในการบรรยายข้อมูล หรือการสร้างฟังก์ชันในการวิเคราะห์ข้อมูล

6. เลือก algorithm(s) ของเหมืองข้อมูลที่จะใช้ในการแก้ปัญหาหรือตอบคำถามที่กำหนดไว้

7. ทำเหมืองข้อมูล

8. การแปลความหมายของเหมืองข้อมูล

9. นำองค์ความรู้ที่ได้ไปใช้

3.3.2 ตัวแบบการดำเนินการ KDD (KDD Process Model) จะมีลักษณะที่คล้ายกับตัวแบบใน 3.3.1 แต่จะมีการลดขั้นตอนให้เหลือเพียง 7 ขั้นตอน (Roiger and Geatz 2003 : 148 -163)คือ

1) กำหนดลักษณะของจุดมุ่งหมาย(Goals identification) วัตถุประสงค์ของขั้นตอนนี้คือ การกำหนดปัญหาหรือคำถามหรือจุดมุ่งหมายที่จะดำเนินการให้ชัดเจน ตัววัดความสำเร็จในการดำเนินการ ความเป็นไปได้ในการดำเนินการ ต้นทุนของการดำเนินการ การเลือกเครื่องที่ใช้ในการทำเหมืองข้อมูล การรับคำปรึกษาของผู้เชี่ยวชาญ การวางแผนการจัดการทรัพยากรมนุษย์และทรัพยากรขององค์กร รวมทั้งการวางแผนการบำรุงรักษาและการปรับปรุงระบบหลังการดำเนินงานเสร็จแล้ว โดยเป็นการวางแผนระยะยาว

2) การสร้างเซตข้อมูลป้าหมาย (Creating a target data set) เป็นการกำหนดตัวแปรที่จะใช้จากแหล่งข้อมูล โดยแหล่งข้อมูลนั้นอาจได้มาจากคลังข้อมูล ฐานข้อมูลธุรกรรม หรือมาจากแฟ้มงานต่างๆ เช่น แฟ้มงานสปรีตชีต(Spreadsheet)

3) การเตรียมข้อมูล(Data preprocessing) เป็นการทำความสะอาดข้อมูล (Data cleaning) ตรวจสอบข้อมูลให้มีความถูกต้อง พร้อมทั้งแจ้งเตือนข้อมูลที่ไม่ถูกต้องให้ทราบก่อนที่ข้อมูลจะถูกนำไปใช้

4) การแปลงข้อมูล(Data transformation) จะเป็นการทำให้ข้อมูลอยู่รูปแบบตามความจำเป็นต่างๆ ซึ่งมีหลายเหตุผลด้วยกัน จะยกตัวอย่างการแปลงข้อมูลที่เราคุ้นเคยกันเช่น

- Data normalization เป็นการแปลงข้อมูลโดยจะเปลี่ยนค่าข้อมูลให้อยู่ในรูปมาตรฐานอันดับที่กำหนด เนื่องจากบางครั้งอัลกอริธึมที่ใช้ในการทำเหมืองข้อมูลมีค่าอยู่ในช่วงที่กำหนด เช่นการใช้อัลกอริธึม Neural networks ในการจำแนกซึ่งค่าเสกลของตัวเลขที่ใช้วัดจะอยู่ในช่วง 0 และ 1 วิธีการ Data normalization มีอยู่หลายวิธี เช่น

Decimal scaling เป็นการหารข้อมูลด้วยตัวเลขโดยส่วนมากตัวเลขที่เรานำมาหารนั้นจะใช้เลขสิบยกกำลัง เช่นถ้าเรารู้ว่าข้อมูลมากและน้อยสุดอยู่ในช่วง -1000 ถึง 1000 ถ้าเราต้องการให้ข้อมูลอยู่ในช่วง -1 ถึง 1 เราจะนำ 1000 มาหาร

Min-Max normalization เป็นเทคนิคที่ต้องรู้ค่าสูงสุดและค่าต่ำสุดของข้อมูล จะทำให้ข้อมูลที่ได้นั้นอยู่ในช่วง 0 และ 1

$$newValue = \frac{originalValue - oldMax}{oldMax - oldMin}$$

การทำให้อยู่ให้รูปคะแนนมาตรฐาน Z (Z-scores) โดยต้องทราบค่าเฉลี่ย (μ) และค่าส่วนเบี่ยงเบนมาตรฐาน(σ)

$$newValue = \frac{originalValue - \mu}{\sigma}$$

- การเปลี่ยนชนิดของข้อมูล (Data type conversion) เทคนิคเหมืองข้อมูลบางเทคนิคไม่สามารถวิเคราะห์ข้อมูลเป็นข้อมูลเชิงกลุ่มได้ จึงต้องทำการแปลงข้อมูลเชิงกลุ่มให้เป็นตัวเลขก่อนการวิเคราะห์

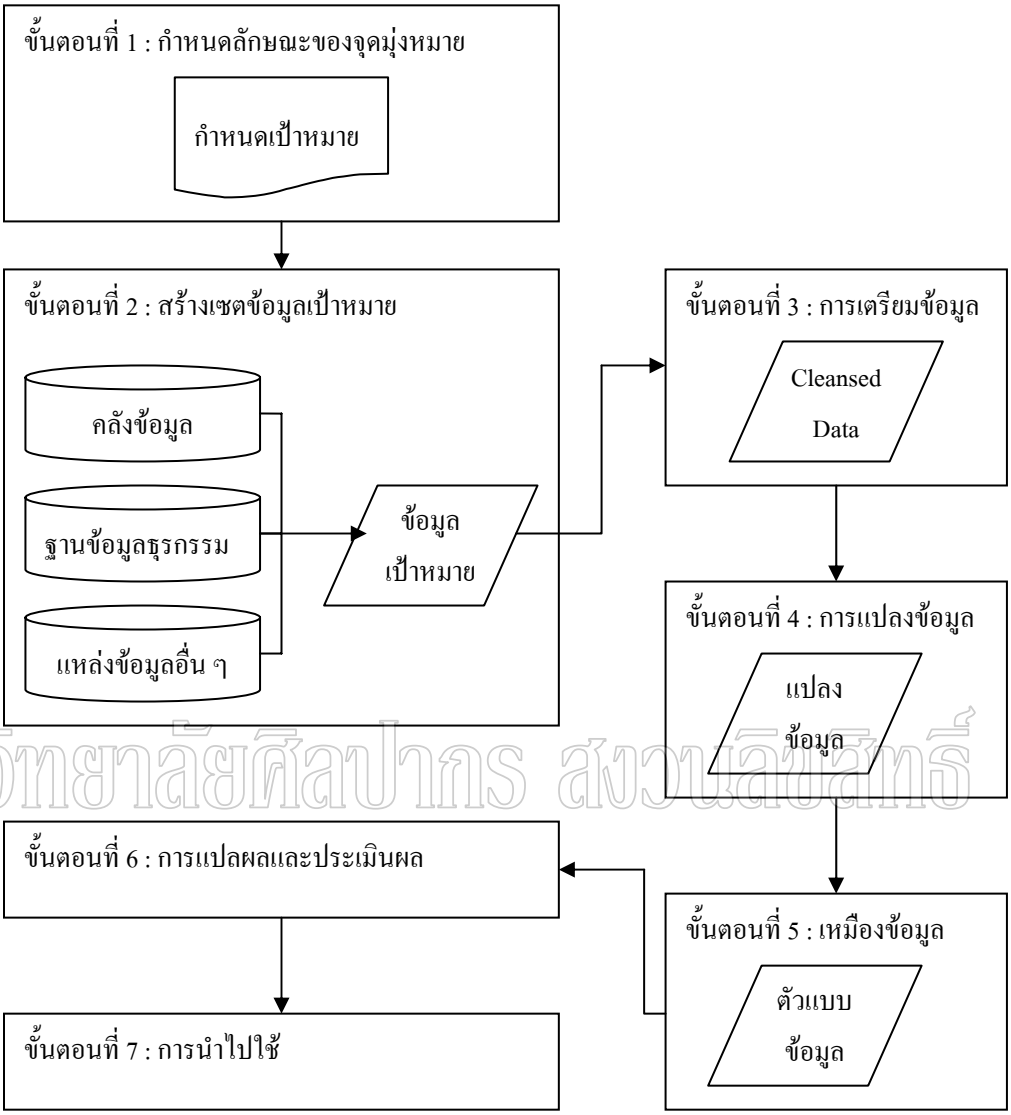
5) การทำเหมืองข้อมูล (Data mining) เป็นการเลือกอัลกอริธึมในการทำเหมืองข้อมูลบางครั้งสามารถเลือกอัลกอริธึมได้หลายวิธี กำหนดข้อมูลว่าจะใช้ข้อมูลใดเป็นข้อมูลที่สร้างและข้อมูลใดใช้ในการทดสอบผลลัพธ์ รวมทั้งการวิเคราะห์ข้อมูล

6) การแปลผลและการประเมินผล(Interpretation and evaluation) เป็นการพิจารณาผลลัพธ์จากขั้นตอนที่ 5 ว่าจะสามารถตอบคำถามหรือแก้ปัญหาจากขั้นตอนที่ 1 หรือไม่ ตัดสินใจว่าจะกระทำในขั้นตอนที่ 5 ซ้ำหรือไม่ และรวมถึงการแปลผลไปให้ยังผู้ใช้ข้อมูลเข้าใจ

7) การนำไปใช้(Taking action) เมื่อลงความเห็นว่าจะนำองค์ความรู้ที่ได้ไปใช้องค์ความรู้นั้นจะถูกรวมเข้าเข้ากับระบบที่ใช้อยู่เช่น

- การสร้างระบบรายงานเกี่ยวกับองค์ความรู้ที่ได้

- การทำระบบจดหมายโดยส่งรายการส่งเสริมการขายให้กับลูกค้าตามกลุ่ม



ภาพที่ 16 Seven-step KDD Process model

ที่มา : Richard J. Roiger and Michael W. Geatz , Data Mining : A Tutorial – Based Primer (Minnesota : Pearson Education Inc, 2003) ,149.

3.3.3 CRISP-DM(Cross-Industry Standard Process for Data Mining) ในปี ค.ศ. 1996 ได้มีความร่วมมือของ DaimlerChrysler(Daimler-Benz) SPSS(ISL) และ NCR เพื่อพัฒนาเครื่องมือในการทำเหมืองข้อมูล และในกลางปี ค.ศ. 1999 จึงได้นำเสนอตัวแบบ CRISP-DM โดยตัวแบบ CRISP-DM จะดำเนินการโครงการเหมืองข้อมูลในลักษณะวงจรชีวิต โดยมี 6 ระยะด้วยกัน(Larose 2005 : 5)คือ

ระยะที่ 1 การทำความเข้าใจธุรกิจ (Business understanding phase) เป็นศูนย์กลางในการพิจารณาวัตถุประสงค์และความต้องการของธุรกิจ เป็นการเริ่มต้นกำหนดปัญหาที่จะใช้เหมืองข้อมูลและเป็นการเริ่มต้นการวางแผนในการพัฒนา

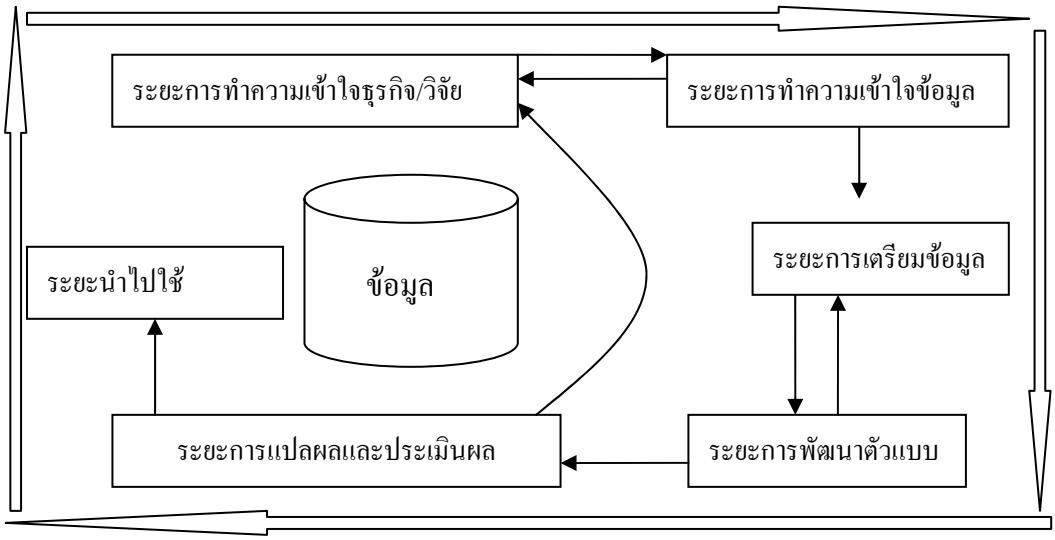
ระยะที่ 2 การทำความเข้าใจข้อมูล (Data undertanding phase) เป็นการพิจารณาชุดข้อมูลที่จะใช้และการตั้งสมมุติฐานที่จะใช้ในการแก้ปัญหา

ระยะที่ 3 การเตรียมข้อมูล (Data preparation phase) เป็นการเลือกแอททริบิวท์ เร็คคอร์ด และตาราง ที่จะใช้ในการวิเคราะห์ รวบรวมข้อมูล รวมทั้งการทำความสะอาดข้อมูล(data clean) สำหรับข้อมูลที่เลือกใช้ในการทำเหมืองข้อมูล

ระยะที่ 4 พัฒนาตัวแบบ(Modeling phase) ระยะนี้จะเป็นการเลือกและประยุกต์เทคนิคที่จะใช้ทำเหมืองข้อมูล

ระยะที่ 5 การแปลผลและการประเมินผล(Evaluation phase) เป็นการวิเคราะห์ผลลัพธ์ที่ได้ว่าผลลัพธ์เป็นไปตามวัตถุประสงค์ที่กำหนดไว้หรือไม่ และกำหนดคุณสมบัติที่จะใช้ของตัวแบบที่บรรลุผลตามวัตถุประสงค์

ระยะที่ 6 การนำไปใช้(Deployment phase) เมื่อตัวแบบของเหมืองข้อมูลเป็นไปตามวัตถุประสงค์แล้ว ระยะที่ 6 จะเป็นการวางแผนและการนำตัวแบบเหมืองข้อมูลไปประยุกต์งาน



ภาพที่ 17 ตัวแบบ CRISP-DM

ที่มา : Daniel T. Larose , Discovering Knowledge in Data :an Introduction to Data Mining (New Jersey : John Wiley & Sons Inc, 2005), 6 .

เมื่อเปรียบเทียบตัวแบบการดำเนินการ KDD ใน 3.3.2 และตัวแบบ CRISP-DM ใน 3.3.3 โดย ระยะที่ 1 และ 2 จะเทียบเท่ากับขั้นตอนที่ 1 ของ ตัวแบบการดำเนินการ KDD และระยะที่ 3 จะรวมขั้นตอนที่ 2 ,3 และ 4 ของตัวแบบการดำเนินการ KDD และระยะที่ 4 ,5 และ 6 จะเหมือนกันกับขั้นตอนที่ 5 ,6 และ 7 ของตัวแบบการดำเนินการ KDD

3.4 เทคนิคเหมืองข้อมูล(Data mining techniques)

จากนิยามของเมืองข้อมูลดังนั้นเทคนิคเหมืองข้อมูลจึงมีวิธีการมากมาย เช่นวิธีทางสถิติ วิธีทางคณิตศาสตร์ วิธีทางคอมพิวเตอร์ Machine learning เป็นต้น โดยเหมืองข้อมูลที่ได้รับความนิยมสามารถแบ่งตามลักษณะการใช้ในการแก้ปัญหาได้ดังนี้

3.4.1 การอธิบายข้อมูล (Description) จะเป็นการวิเคราะห์ข้อมูลที่ไม่ซับซ้อนมากนักมักจะเป็นการอธิบายรูปแบบและแนวโน้มที่อยู่ในข้อมูลหรือการสรุปข้อมูล เช่น การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) การทำ Data visualization การทดสอบสมมุติฐาน

3.4.2 การจำแนกกลุ่ม(Classification) เป็นการจำแนกของวัตถุที่ต้องการโดยอาศัยลักษณะที่คล้ายคลึงกันหรือแตกต่างกัน เป็นการจัดวัตถุที่เข้ามาใหม่เข้ากลุ่มที่จัดไว้แล้วให้เข้ากลุ่มได้ถูกต้อง โดยเราต้องรู้ว่าวัตถุที่ต้องการจัดเข้ากลุ่มนั้นมีจำนวนกลุ่มที่แน่นอน โดยตัวแปรตามจะเป็นตัวแปรเชิงกลุ่ม การจำแนกกลุ่มจะเน้นการสร้างตัวแบบที่กำหนดว่าวัตถุที่เข้ามาใหม่นั้นจะถูกจัดเข้ากลุ่มใดกลุ่มหนึ่งที่กำหนด เช่น การจำแนกลูกค้าที่ค้างชำระจะสามารถชำระหนี้ได้หรือไม่ หรือการจำแนก

ลูกค้าของการให้เครดิตว่าจะมีความเสี่ยงมาก ปานกลาง หรือน้อย เทคนิคที่ใช้ เช่น Neural network ,Discriminant analysis ,Decision tree เป็นต้น

3.4.3 การรวมกลุ่ม(Clustering) เป็นการรวมวัตถุที่คล้ายคลึงกันไว้ในกลุ่มเดียวกัน โดยจะไม่มีข้อสมมุติเกี่ยวกับจำนวนกลุ่มว่ามีกี่กลุ่ม แตกต่างจากการจำแนกกลุ่มตรงที่การจำแนกกลุ่มเป็นการจัดวัตถุใหม่เข้ากลุ่มใดกลุ่มหนึ่งจากกลุ่มที่มีอยู่ การรวมกลุ่มจะไม่มีตัวแปรตามแต่จะมีแต่ตัวแปรอิสระที่ใช้วัดความคล้ายคลึงหรือใช้ในการคำนวณความคล้ายคลึง เช่น การรวมกลุ่มของลูกค้าที่มีพฤติกรรมการซื้อที่มีลักษณะคล้ายคลึงกัน เทคนิคที่ใช้ เช่น K-Means เทคนิครวมกลุ่มแบบมีลำดับชั้น เป็นต้น

3.4.4 การประมาณค่า(Estimation) จุดประสงค์หลักในการประมาณค่านั้นจะเน้นการกำหนดค่าของตัวแปรตามที่เราไม่ทราบค่าจากตัวแปรอิสระต่างๆ ซึ่งตัวแปรตามนั้นจะมีค่าเป็นตัวเลขหรือมีลักษณะเป็นข้อมูลชนิดต่อเนื่อง เช่น การประมาณรายได้ของพนักงานแต่ละคน

3.4.5 การทำนาย(Prediction) จะมีลักษณะคล้ายกับการประมาณค่า แต่ตัวแบบในการทำนายจะมุ่งเน้นเป็นการศึกษาพฤติกรรมในอนาคตมากกว่าในปัจจุบัน โดยตัวแปรตามที่จะทำนายนั้นจะเป็นข้อมูลชนิดต่อเนื่องหรือไม่ต่อเนื่องก็ได้ เช่น การทำนายยอดขายของบริษัทในเดือนหน้า

3.4.6 Affinity grouping หรือ Association rule เป็นการหาความสัมพันธ์หรือความเกี่ยวเนื่องของข้อมูลโดยอาศัยหลักของกฎ ซึ่งจะอยู่ในรูปแบบ “ถ้า *สิ่งที่เกิดขึ้น* แล้ว *ผลที่จะตามมา* (if *antecedent* then *consequent*) “ (Larose 2005 : 17) เช่น การวิเคราะห์ Market basket analysis ซึ่งเป็นการวิเคราะห์พฤติกรรมการซื้อสินค้าซึ่งอาศัยวิธีการของ Association rule เป็นหลัก ตัวอย่างเช่น ถ้ามีสินค้าอยู่ 4 ประเภทคือ นม เนย ขนมปัง และ ไข่ เราสามารถหากฎที่เป็นไปได้ดังนี้

1. ถ้าลูกค้าซื้อนมแล้วซื้อขนมปังด้วย
2. ถ้าลูกค้าซื้อขนมปังแล้วซื้อนมด้วย
3. ถ้าลูกค้าซื้อนมและไข่แล้วซื้อเนย และขนมปังด้วย
4. ถ้าลูกค้าซื้อนม เนย และไข่แล้วซื้อขนมปังด้วย

นอกจากนี้เราอาจจะแบ่งประเภทของเหมืองข้อมูลออกเป็น 2 ประเภทคือ Unsupervised และ Supervised โดยวิธีแบบ Unsupervised จะไม่มีตัวแปรตาม(Dependent variable) แต่จะเป็นการหารูปแบบ โครงสร้างหรือความสัมพันธ์ของข้อมูลจากตัวแปรอิสระ(Independent variable) เช่น การรวมกลุ่ม แต่ วิธีแบบ Supervised จะเป็นอธิบายรายละเอียด การหาหรือกำหนดค่าให้กับตัวแปรตามจากตัวแปรอิสระ(Roiger and Geatz 2003 : 34) เช่น การพยากรณ์ เป็นต้น

ดังที่กล่าวมาแล้วว่าการทำเหมืองข้อมูลมีเทคนิคหรือวิธีการมากมาย การเลือกใช้เทคนิคต่างๆ นั้นขึ้นอยู่ว่าจะใช้ในการแก้ปัญหาหรือการหาคำตอบที่เรากำหนดขึ้นได้อย่างไร บางครั้ง

เทคนิคเดียวกันแต่เราสามารถใช้แก้ปัญหาที่แตกต่างกันได้ และในบางครั้งปัญหาเดียวกันสามารถใช้เทคนิคที่แตกต่างกันในการแก้ปัญหา ต่อไปจะนำเสนอเทคนิคบางเทคนิคที่ใช้ในเหมืองข้อมูล

1. Data visualization

จะนำเสนอข้อมูลในลักษณะของแผนภาพ รูปภาพ หรือกราฟแทนการอธิบายข้อมูลโดยข้อความ ซึ่งอาจจะแสดงข้อมูลในรูปแบบสองมิติหรือสามมิติ

2. สหสัมพันธ์เชิงเส้น(Linear correlation)

เป็นการวิเคราะห์ข้อมูลเชิงสถิติภายใต้ตัวแปร 2 ตัว(X,Y) โดยที่ตัวแปรวัดมาจากหน่วยสังเกต หน่วยทดลอง หน่วยประชากร หรือ หน่วยตัวอย่าง เช่น น้ำหนัก (X) กับ ส่วนสูง (Y) การวิเคราะห์จะเป็นการหาความสัมพันธ์เชิงเส้นระหว่างสองตัวแปร ก่อนอื่นจะต้องกล่าวถึงความแปรปรวนร่วมระหว่าง X และ Y แทนโดยสัญลักษณ์ Cov(X,Y) หรือ σ_{XY} โดย

$$Cov(X,Y) = \sigma_{XY} = E(X-\mu_X)(Y-\mu_Y)$$

เมื่อ μ_X, μ_Y เป็นค่าเฉลี่ยของ X และ Y ในประชากรตามลำดับ

ส่วนความแปรปรวนของ X คือ

$$Cov(X,X) = \sigma_{XX} = \sigma_X^2 = E(X-\mu_X)(X-\mu_X) = E(X-\mu_X)^2$$

และความแปรปรวนของ Y จะสามารถนิยามเช่นเดียวกับความแปรปรวนของ X

ความแปรปรวนร่วมจะเป็นตัวชี้วัดความสัมพันธ์ระหว่าง X และ Y โดยถ้า Cov(X,Y) > 0 จะมีความสัมพันธ์เชิงบวก Cov(X,Y) < 0 จะมีความสัมพันธ์เชิงลบ และ Cov(X,Y) = 0 จะไม่มีความสัมพันธ์ แต่ความแปรปรวนร่วมจะเป็นตัววัดที่ขึ้นอยู่กับหน่วยวัด(Scale dependence)(วีรานันท์ พงศาภักดี ม.ป.ป.: 9-2) ดังนั้นการที่จะให้หน่วยวัดหมดไปนั้น เราจะนำเสนอความสัมพันธ์ในเทอมสัมประสิทธิ์สหสัมพันธ์(Coefficient of correlation) แทนการนำเสนอความแปรปรวนร่วม โดยสัมประสิทธิ์สหสัมพันธ์(ρ) สามารถคำนวณได้ดังนี้

$$\rho = \frac{Cov(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

$$= \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

ค่าสัมประสิทธิ์สหสัมพันธ์จะไม่ขึ้นอยู่กับหน่วยวัด และมีคุณสมบัติดังนี้

1. จะมีค่าอยู่ระหว่าง -1 ถึง 1 $(-1 \le \rho \le 1)$ ถ้า ρ มีค่าเข้าใกล้ -1 หรือ 1 แสดงว่าข้อมูลมี ความสัมพันธ์กันมาก ถ้า ρ มีค่าเข้าใกล้ 0 แสดงว่าข้อมูลมี ความสัมพันธ์กันน้อย และ $\rho = 0$ จะแสดงว่าข้อมูลไม่มีความสัมพันธ์กัน
2. $\rho = 1$ เมื่อ $Y = \alpha + \beta X$ โดย $\beta > 1$
   $\rho = -1$ เมื่อ $Y = \alpha + \beta X$ โดย $\beta < 1$

ตัวประมาณของสัมประสิทธิ์สหสัมพันธ์(r) เมื่อสุ่มข้อมูลขนาด n จากตัวแปรค่าสังเกต (X,Y) โดยมีค่าเป็น $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ ซึ่งตัวประมาณของความแปรปรวนร่วมของตัวอย่าง (S_{xy}) โดยที่

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$= \frac{1}{n-1}(\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n})$$

จะเป็นตัวประมาณของ σ_{xy}

ดังนั้นตัวประมาณของสัมประสิทธิ์สหสัมพันธ์ (ρ) จึงใช้เป็นสัมประสิทธิ์สหสัมพันธ์ตัวอย่าง (r) ดังนี้

$$r = \frac{S_{xy}}{S_x S_y}$$

เมื่อ S_x คือ ส่วนเบี่ยงเบนมาตรฐาน(Standard deviation)ของตัวอย่างสุ่มของตัวแปร X

S_y คือ ส่วนเบี่ยงเบนมาตรฐาน(Standard deviation)ของตัวอย่างสุ่มของตัวแปร Y

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2]^{\frac{1}{2}}}$$

ซึ่งตัวประมาณ r นี้ เรียกว่า สัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน(Pearson's correlation coefficient)

3. การวิเคราะห์การถดถอยเชิงเส้น(Linear regression analysis)

การวิเคราะห์การถดถอยเป็นการวิธีทางสถิติที่ใช้ในการทำนายหรือพยากรณ์ของค่าตัวแปรตาม(Dependent variable)จากตัวแปรอิสระ(Independent variable)โดยที่ตัวแปรอิสระอาจจะมีตัวแปรอิสระเพียงตัวแปรเดียวหรือมากกว่าหนึ่งตัวแปรอิสระ โดยตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์เชิงเส้น

- การวิเคราะห์ถดถอยอย่างง่าย(Simple regression analysis) เป็นการวิเคราะห์ระหว่างตัวแปรเชิงปริมาณสองตัวซึ่งตัวหนึ่งเป็นตัวแปรตามและอีกตัวเป็นตัวแปรอิสระ เช่น ตัวแปรX แทนรายได้ของประชากรในปี 2547-2549 และตัวแปร Y แทนรายจ่ายของประชากรในปี 2547-2549 โดยตัวแบบของการวิเคราะห์ถดถอยอย่างง่าย ดังนี้

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

เมื่อ Y เป็นตัวแปรตาม

X เป็นตัวแปรอิสระ

β_0, β_1 เป็นพารามิเตอร์ที่ไม่ทราบค่า

ε เป็นความคลาดเคลื่อนสุ่ม(Random error)โดยแต่ละค่าของ X นั้น ε มีการแจกแจงแบบปกติและมีค่าเฉลี่ยเป็น 0 ความแปรปรวนเท่ากับ σ^2 , $\varepsilon \sim N(0, \sigma^2)$

โดยจะต้องมีข้อตกลงเบื้องต้น(Assumption)ดังนี้

1. $\mu_{Y|X} = E(Y \mid X) = \beta_0 + \beta_1 X$ หมายความว่า ค่าเฉลี่ยของ Y มีรูปแบบเชิงเส้น หรือการถดถอยของ Y บน X เป็นแบบเชิงเส้นตรงภายใต้พารามิเตอร์ β_0, β_1

2. $Var(Y \mid X) = \sigma^2_{Y|X} = \sigma^2$ หมายความว่า ความแปรปรวนของการถดถอยมีค่าคงที่และไม่เปลี่ยนไปตาม X

3. แต่ละ X ใดๆ Y เป็นตัวแปรสุ่มที่มีค่าเฉลี่ยและความแปรปรวนที่ขึ้นอยู่กับ X ตามข้อ 1,2 ตามลำดับ

4. ค่าของ Y เป็นอิสระต่อกัน

5. Y มีการแจกแจงแบบปกติโดยมีค่าเฉลี่ยและความแปรปรวนตามข้อ 1,2 ตามลำดับ $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

การประมาณ β_0, β_1 โดยวิธีกำลังสองน้อยที่สุด(Method of least squares) เป็นวิธีที่นิยมมากที่สุดโดยจากข้อมูลของตัวแปร X และ Y มีค่าสังเกต $(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)$ วิธีกำลังสองน้อยที่สุดจะสามารถประมาณค่าของ β_0, β_1 ได้ดังนี้

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

ดังนั้นตัวแบบการถดถอยของ Y บน X ที่ประมาณค่า β_0, β_1 ด้วย $\hat{\beta}_0, \hat{\beta}_1$ จะมีรูปแบบดังนี้

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ(Multiple linear regression analysis) เป็นการพยากรณ์ตัวแปรตามหนึ่งตัวแปรโดยใช้ตัวแปรอิสระหลายตัว ซึ่งจะทำให้การพยากรณ์ทำได้ดียิ่งขึ้น สมมุติให้ Y เป็นตัวแปรตาม และ X เป็นตัวแปรอิสระ p ตัว ดังนั้นตัวแบบของการถดถอยของ Y บน $X_1, X_2, ..., X_p$ คือ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

เมื่อ Y เป็นตัวแปรตาม

$X_1, X_2, ..., X_p$ เป็นตัวแปรอิสระ

ε เป็นความคลาดเคลื่อนสุ่ม(Random error)

โดยเมื่อค่าสังเกตของตัวแปร Y เป็นค่าที่เป็นอิสระกัน n ค่า พร้อมทั้งค่าของตัวแปร X ทั้ง p ตัวที่สัมพันธ์กับ Y ดังนั้นตัวแบบที่สมบูรณ์คือ

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \varepsilon_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \varepsilon_n
\end{aligned}$$

โดยข้อตกลงเบื้องต้นของ ε มีดังนี้

1. ε_i เป็นตัวแปรสุ่มมีค่าเฉลี่ยเท่ากับ0 มีค่าความแปรปรวนคงที่ นั่นคือ $E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2$ ทุกค่าของ i

2. ความแปรปรวนร่วมของ ε_i กับ ε_j เท่ากับ 0 หรือ ε_i กับ ε_j ไม่มีสหสัมพันธ์ เมื่อ $i \neq j$, $Cov(\varepsilon_i, \varepsilon_j) = 0$

3. ε_i มีการแจกแจงแบบปกติมีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ $\sigma^2, \varepsilon_i \sim N(0, \sigma^2)$ เมื่อ i = 1,2,3,…,n

โดยการถดถอยเชิงเส้นพหุคูณสามารถเขียนอยู่ในรูปของเมตริกซ์ได้ดังนี้

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

หรือ

$$\underset{(n\times 1)}{Y} = \underset{(n\times(r+1))}{X} \underset{((r+1)\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon}$$

เมื่อ Y เป็นเวกเตอร์ของตัวแปรตาม

X เป็นเวกเตอร์ของตัวแปรอิสระ

β เป็นเวกเตอร์พารามิเตอร์

ε เป็นเวกเตอร์ความคลาดเคลื่อนสุ่ม โดย $E(\varepsilon)=0, Cov(\varepsilon)=E(\varepsilon\varepsilon')=\sigma^2 I$, ε มีการแจกแจงแบบพหุปกติ , $\varepsilon \sim N(0,\sigma^2 I)$

การประมาณค่าของ β โดยวิธีกำลังสองน้อยที่สุด โดยเราให้ b เป็นเวกเตอร์ประมาณค่าของ β ซึ่ง $b=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$ และให้ e เป็นเวกเตอร์ประมาณค่าของ ε (เวกเตอร์ส่วนเหลือ) ซึ่ง $e=\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$ ดังนั้นตัวประมาณ $b=(X'X)^{-1}X'y$ และมีค่าความแปรปรวน $V(b)=\sigma^2(X'X)^{-1}$ ซึ่งเราสามารถประมาณค่าของ σ^2 ได้จาก $s^2=\frac{\sum_{i=1}^{n} e_i^2}{n-p-1}=\frac{e'e}{n-p-1}$ จากการประมาณค่าของ β โดยวิธีกำลังสองน้อยที่สุด เราสามารถเขียนสมการที่ใช้ในการทำนายหรือพยากรณ์ค่า $\hat{Y}$ ได้คือ

$$\hat{Y}=Xb$$

และมีค่าความคลาดเคลื่อน หรือส่วนเหลือ (Residual) คือ

$$e=Y-\hat{Y}$$

4. Association rule

จากที่กล่าวมาแล้วว่า Association rule เป็นการหาความสัมพันธ์หรือความเกี่ยวเนื่องของข้อมูลโดยอาศัยหลักของกฎ ซึ่งจะอยู่ในรูปแบบ “ถ้า *สิ่งเกิดที่ขึ้น* แล้ว *ผลที่จะตามมา*” ใน Association rule นั้นค่าที่เป็นไปได้ของแอททริบิวท์จะเป็นลักษณะเพียงสองลักษณะซึ่งจะแทนด้วย 0 หรือ 1 , ใช่หรือ ไม่ ดังนั้นจำนวนที่เป็นไปได้ของแอททริบิวท์ของ Association rule ในกรณีที่มีแอททริบิวท์จำนวน k แอททริบิวท์ จะมีจำนวน Association rule ที่เป็นไปได้คือ $k.2^{k-1}$(Larose 2005 : 180-181) ตัวอย่างเช่นถ้าในร้านมีสินค้าที่แตกต่างกัน 100 ชิ้น จะมีกรณีที่เป็นไปได้ใน Association rule คือ $100.2^{99} \approx 6.4\times10^{31}$ และตัววัดของความเกี่ยวเนื่องของกฎคือ Support และ Confidence ตัวอย่างเช่น ในร้านสะดวกซื้อมีลูกค้าที่ซื้อของ 1,000 คนในวันอังคาร โดยมีคนที่ซื้อผ้าอ้อม 200 คน และใน 200 คนนั้นจะมีอยู่ 50 คนที่ซื้อเบียร์ ดังนั้นการทำ Association rule โดย ถ้าซื้อผ้าอ้อมแล้วจะซื้อเบียร์นั้น Support จะเท่ากับ 50/1000 = 5% และ Confidence จะเท่ากับ 50/200 =25% ให้ D เป็นจำนวนTransaction ทั้งหมดที่เกิดขึ้น

โดย Support(s) สำหรับ Association rule เฉพาะกรณีของ $A \Rightarrow B$ คือ สัดส่วนของ Transaction ใน D ที่มีทั้ง A และ B อยู่ (Larose 2005 : 184) นั่นคือ

Support = $P(A \cap B)$ = (จำนวนของTransactionที่มีAและ B)/(จำนวนTransactionทั้งหมด)

และ Confidence (c) ของ Association rule ของ $A \Rightarrow B$ ซึ่งเป็นตัววัดความถูกต้องหรือความแน่นอนของกฎ เป็นการกำหนดเปอร์เซ็นต์ของจำนวน Transaction ใน D โดย Transaction ของ A ที่มี B อยู่ด้วย(Larose 2005 : 184) นั่นคือ

Confidence = $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$ =(จำนวนของTransactionที่มีAและ B)/(จำนวน Transactionที่มี A)

การวิเคราะห์ก่อนที่เราจะเสนอกฎนั้นควรจะมีค่าของ Support หรือ Confidence ที่สูง หรือให้ทั้งสองค่ามีค่าสูง โดยทั่วไปแล้วกฎที่แข็ง(Strong rule)นั้นจะกำหนดค่าขั้นต่ำของ Support และ Confidence ไว้ในระดับที่สูงเช่น ในการวิเคราะห์สนใจจะค้นหาความเกี่ยวพันธ์ในการซื้อของในซูปเปอร์มาเกตซึ่งกำหนดขั้นต่ำของ Support ที่ระดับ 20 % และ Confidence ขั้นต่ำที่ระดับ 70%

ชุดรายการคือเซตของรายการ(Itemset)ที่อยู่ใน I และ k ชุดรายการคือรายการที่มี k รายการ เช่น {ถั่ว,แตง} คือ 2 ชุดรายการ ,และ {บร็อคโคลี่,พริกหวาน,ข้าวโพด} คือ 3 ชุดรายการ ซึ่งผักแต่ละชนิดจะอยู่ในเซต I ความถี่ชุดรายการ(Itemset frequency)คือจำนวนของTransaction ที่มีเฉพาะชุดรายการนั้น ชุดรายการที่เกิดบ่อย(Frequent itemset)คือ ชุดรายการที่เกิดรายการที่กำหนดมีความถี่ของชุดรายการมากกว่าหรือเท่ากับ c เมื่อ c คือค่าคงที่ ตัวอย่างเช่น กำหนดให้ c = 4 แล้วชุดรายการที่เกิดมากกว่าหรือเท่ากับ 4 ครั้ง จะกล่าวว่าเกิดขึ้นบ่อย เราจะกำหนดว่าเซตที่เกิดบ่อย k-ชุดรายการคือ F_k

A priori algorithm เป็นวิธีหนึ่งของ Association rule โดยคุณสมบัติของ A priori คือ “ถ้าชุดรายการ Z ไม่เกิดบ่อยแล้ว สำหรับทุกรายการ A , $Z \cup A$ จะไม่เกิดบ่อย ” (Larose 2005 : 185) ในการทำ Association rule มีขั้นตอนการดำเนินการอยู่ 2 ขั้นตอน (Larose 2005 : 184) คือ

1. ทำการค้นหาความถี่ของชุดรายการทั้งหมด นั่นคือ หาชุดรายการที่มีความถี่และไม่เป็นเซตว่าง

2. จากชุดรายการที่ได้ ทำการสร้าง Association rule โดยกำหนดเงื่อนไขขั้นต่ำที่เพียงพอของ Support และ Confidence

ตัวอย่างการทำ Association rule

จากข้อมูลการซื้อผักในร้านค้าแห่งหนึ่งซึ่งมีผักขายจำนวน 7 ชนิดคือ หน่อไม้ฝรั่ง , ถั่ว , บร็อคโคลี่,ข้าวโพด,พริกหวาน,แตง,มะเขือเทศ(Larose 2005 : 182-188) โดยมีข้อมูลการซื้อดังนี้

ตารางที่ 13 ตัวอย่างรายการซื้อสินค้า

Transaction	รายการซื้อ
1	บร็อคโคลี่,พริกหวาน,ข้าวโพด
2	หน่อไม้ฝรั่ง,แตง,ข้าวโพด
3	ข้าวโพด,มะเขือเทศ,ถั่ว,แตง
4	พริกหวาน,ข้าวโพด,มะเขือเทศ,ถั่ว
5	ถั่ว,หน่อไม้ฝรั่ง,บร็อคโคลี่
6	แตง,หน่อไม้ฝรั่ง,ถั่ว,มะเขือเทศ
7	มะเขือเทศ,ข้าวโพด
8	บร็อคโคลี่,มะเขือเทศ,พริกหวาน
9	แตง,หน่อไม้ฝรั่ง,ถั่ว
10	ถั่ว,ข้าวโพด
11	พริกหวาน,บร็อคโคลี่,ถั่ว,แตง
12	หน่อไม้ฝรั่ง,ถั่ว,แตง
13	แตง,ข้าวโพด,หน่อไม้ฝรั่ง,ถั่ว
14	ข้าวโพด,พริกหวาน,มะเขือเทศ,ถั่ว,บร็อคโคลี่

ที่มา : Daniel T. Larose , Discovering Knowledge in Data :an Introduction to Data Mining (New Jersey : John Wiley & Sons Inc ,2005), 182 .

จากข้อมูลดังกล่าว เราสามารถนำข้อมูลมาแจกแจงให้อยู่ในรูปของตารางดังนี้

ตารางที่ 14 ตัวอย่างการแจกแจงข้อมูลการซื้อ

Transaction	หน่อไม้ฝรั่ง	ถั่ว	บร็อคโคลี่	ข้าวโพด	พริกหวาน	แตง	มะเขือเทศ
1	0	0	1	1	1	0	0
2	1	0	0	1	0	1	0
3	0	1	0	1	0	1	1
4	0	1	0	1	1	0	1
5	1	1	1	0	0	0	0
6	1	1	0	0	0	1	1
7	0	0	0	1	0	0	1

ตารางที่ 14 (ต่อ)

Transaction	หน่อไม้ฝรั่ง	ถั่ว	บร็อคโคลี่	ข้าวโพด	พริกหวาน	แตง	มะเขือเทศ
8	0	0	1	0	1	0	1
9	1	1	0	0	0	1	0
10	0	1	0	1	0	0	0
11	0	1	1	0	1	1	0
12	1	1	0	0	0	1	0
13	1	1	0	1	0	1	0
14	0	1	1	1	1	0	1

กำหนดให้ c = 4 ชุดรายการที่เกิดบ่อยถ้าเกิดรายการมากกว่าหรือเท่ากับ 4 หา F_1 ที่เกิดบ่อย 1-ชุดรายการ จากข้อมูลผักทุกชนิดมีจำนวน Transaction เกิดขึ้นมากกว่า c ดังนั้น F_1 = {หน่อไม้ฝรั่ง,ถั่ว,บร็อคโคลี่,ข้าวโพด,พริกหวาน,แตง,มะเขือเทศ}

ต่อไปหา F_2 ที่เกิดบ่อย 2-ชุดรายการ โดยทั่วไปแล้วการหา F_k จาก A priori algorithm ขั้นแรกต้องสร้างเซต C_k ซึ่งมี k-ชุดรายการโดยใช้การรวมกันของ F_{k-1} แล้วตัด C_k ออก โดยใช้คุณสมบัติ A priori ชุดรายการใน C_k ที่ไม่ถูกตัดออกจะอยู่ในรูป F_k

ตารางที่ 15 สรุปจำนวนของ 2-ชุดรายการจากตัวอย่างการซื้อสินค้า

Combination	Count	Combination	Count
หน่อไม้ฝรั่ง,ถั่ว	5	บร็อคโคลี่,ข้าวโพด	2
หน่อไม้ฝรั่ง,บร็อคโคลี่	1	บร็อคโคลี่,พริกหวาน	4
หน่อไม้ฝรั่ง,ข้าวโพด	2	บร็อคโคลี่,แตง	1
หน่อไม้ฝรั่ง,พริกหวาน	0	บร็อคโคลี่,มะเขือเทศ	2
หน่อไม้ฝรั่ง,แตง	5	ข้าวโพด,พริกหวาน	3
หน่อไม้ฝรั่ง,มะเขือเทศ	1	ข้าวโพดมแตง	3
ถั่ว,บร็อคโคลี่	3	ข้าวโพด,มะเขือเทศ	4
ถั่ว,ข้าวโพด	5	พริกหวาน,แตง	1
ถั่ว,พริกหวาน	3	พริกหวาน,มะเขือเทศ	3
ถั่ว,แตง	6	แตง,มะเขือเทศ	2
ถั่ว,มะเขือเทศ	4		

เนื่องจาก c = 4 เราจะมี F_2 ={{หน่อไม้ฝรั่ง,ถั่ว},{หน่อไม้ฝรั่ง,แตง},{ถั่ว,ข้าวโพด},{ถั่ว,แตง},{ถั่ว,มะเขือเทศ},{บร็อคโคลี่,พริกหวาน},{ข้าวโพด,มะเขือเทศ}} ต่อไปเราจะใช้ F_2 ในการสร้าง C_3 ซึ่งสมาชิกเป็น 3-ชุดรายการ โดยการนำสมาชิกของ F_2 มารวมเช่น {หน่อไม้ฝรั่ง,ถั่ว} และ {หน่อไม้ฝรั่ง,แตง} จะเป็น 3-ชุดรายการของ {หน่อไม้ฝรั่ง,ถั่ว,แตง} ทำนองเดียวกัน {ถั่ว,ข้าวโพด} และ {ถั่ว,แตง} จะเป็น 3-ชุดรายการของ {ถั่ว,ข้าวโพด,แตง} ดังนั้น C_3 = {{หน่อไม้ฝรั่ง,ถั่ว,แตง} , {ถั่ว,ข้าวโพด,แตง},{ถั่ว,ข้าวโพด,มะเขือเทศ},{ถั่ว,แตง,มะเขือเทศ}}

แล้วจะทำการตัดสมาชิกของ C_3 โดยใช้คุณสมบัติ A Priori โดยสำหรับแต่ละ s เมื่อ s เป็นสมาชิกใน C_3 ซึ่งจะสร้างสับเซตขนาด k-1 จากการแยก s ซึ่งถ้าซับเซตมีสมาชิกที่มีความถี่น้อยกว่า c จะทำให้ s ถูกตัดออกจาก C_3 ตัวอย่างเช่น ให้ s = {หน่อไม้ฝรั่ง,ถั่ว,แตง} ดังนั้นสับจะมีขนาด k-1 = 2 จะสร้างได้ดังนี้ {หน่อไม้ฝรั่ง,ถั่ว} , {หน่อไม้ฝรั่ง,ถั่ว} และ {ถั่ว,แตง} เมื่อพิจารณาสับเซตแต่ละสับเซตกับตารางที่ 15 พบว่ามีจำนวนมากกว่า c ทุกสมาชิก ดังนั้น s ไม่ถูกตัดออก อย่างไรก็ตามให้ s = {ถั่ว,ข้าวโพด,แตง} เมื่อพิจารณาสับเซต {ข้าวโพด,แตง} มีความถี่ 3 < c ดังนั้น {ถั่ว,ข้าวโพด,แตง} จะถูกตัดออกจาก C_3

เมื่อพิจารณาทุกสมาชิกของ C_3 แล้วจะได้ว่า F_3 ={{หน่อไม้ฝรั่ง,ถั่ว,แตง} } จะทำการหา F_4 แต่เนื่องจากต้องสร้าง C_4 จาก F_3 แต่ F_3 มีสมาชิกเพียงสมาชิกเดียวดังนั้นตั้งแต่ F_4 จะไม่เกิดขึ้นบ่อย ซึ่งจากข้อมูลเราจะสร้าง Association rule และหาค่าของ Support และ Confidence ได้ดังตารางข้างล่าง

ตารางที่ 16 การทำ Association rule : two antecedent จากข้อมูลตัวอย่าง

If antecedent then consequent	Support	Confidence
ถ้า ซื้อหน่อไม้ฝรั่งและถั่ว แล้ว ซื้อแตง	4/14 = 28.6%	4/5 = 80 %
ถ้า ซื้อหน่อไม้ฝรั่งและแตง แล้ว ซื้อถั่ว	4/14 = 28.6%	4/5 = 80 %
ถ้าซื้อ ถั่วและแตง แล้ว ซื้อหน่อไม้ฝรั่ง	4/14 = 28.6%	4/6 = 66.7 %

ตารางที่ 17 การทำ Association rule :one antecedent จากข้อมูลตัวอย่าง

If antecedent then consequent	Support	Confidence
ถ้า ซื้อหน่อไม้ฝรั่ง แล้ว ซื้อ ถั่ว	5/14 = 35.7 %	5/6 = 83.3 %
ถ้า ซื้อถั่ว แล้ว ซื้อ หน่อไม้ฝรั่ง	5/14 = 35.7 %	5/10 = 50 %
ถ้า ซื้อหน่อไม้ฝรั่ง แล้ว ซื้อแตง	5/14 = 35.7 %	5/6 = 83.3 %
ถ้า ซื้อแตง แล้ว ซื้อหน่อไม้ฝรั่ง	5/14 = 35.7 %	5/7 = 71.4 %
ถ้า ซื้อถั่ว แล้ว ซื้อข้าวโพด	5/14 = 35.7 %	5/10 = 50 %
ถ้า ซื้อข้าวโพด แล้ว ซื้อถั่ว	5/14 = 35.7 %	5/8 = 62.5 %
ถ้า ซื้อถั่ว แล้ว ซื้อแตง	6/14 = 42.9 %	6/10 = 60 %
ถ้า ซื้อแตง แล้ว ซื้อถั่ว	6/14 = 42.9 %	6/7 = 85.7 %
ถ้า ซื้อถั่ว แล้ว ซื้อมะเขือเทศ	4/14 = 28.6 %	4/10 = 40 %
ถ้า ซื้อมะเขือเทศ แล้ว ซื้อถั่ว	4/14 = 28.6 %	4/6 = 66.7 %
ถ้า ซื้อ บร็อคโคลี่ แล้ว ซื้อพริกหวาน	4/14 = 28.6 %	4/5 =80 %
ถ้า ซื้อพริกหวาน แล้ว ซื้อ บร็อคโคลี่	4/14 = 28.6 %	4/5 =80 %
ถ้า ซื้อข้าวโพด แล้ว ซื้อมะเขือเทศ	4/14 = 28.6 %	4/8 = 50 %
ถ้า ซื้อมะเขือเทศ แล้ว ซื้อข้าวโพด	4/14 = 28.6 %	4/6 = 66.7 %

เพื่อให้ Association rule ที่สร้างขึ้นมีการใช้งานอย่างมีประโยชน์และมีประสิทธิภาพ จึงทำการกำหนดค่าของ Support หรือ Confidence ให้มีค่าสูง ในตัวอย่างเราจะกำหนดให้ค่าของ confidence มีค่าตั้งแต่ 80 % เป็นต้นไปถึงนำไปใช้ได้ ซึ่งจะแสดงดังตารางข้างล่างและจะเรียงกฎที่ได้ตาม Support × Confidence จากมากไปหาน้อย

ตารางที่ 18 ผลการทำ Association rule จากข้อมูลตัวอย่าง

If antecedent then consequent	Support	Confidence	Support ×Confidence
ถ้า ซื้อแตง แล้ว ซื้อถั่ว	6/14 = 42.9 %	6/7 = 85.7 %	0.3677
ถ้า ซื้อหน่อไม้ฝรั่ง แล้ว ซื้อ ถั่ว	5/14 = 35.7 %	5/6 = 83.3 %	0.2974
ถ้า ซื้อหน่อไม้ฝรั่ง แล้ว ซื้อแตง	5/14 = 35.7 %	5/6 = 83.3 %	0.2974
ถ้า ซื้อ บร็อคโคลี่ แล้ว ซื้อพริกหวาน	4/14 = 28.6 %	4/5 =80 %	0.2288

ตารางที่ 18 (ต่อ)

If antecedent then consequent	Support	Confidence	Support ×Confidence
ถ้า ซื้อพริกหวาน แล้ว ซื้อ บร็อคโคลี่	4/14 = 28.6 %	4/5 =80 %	0.2288
ถ้า ซื้อหน่อไม้ฝรั่งและถั่ว แล้ว ซื้อแตง	4/14 = 28.6%	4/5 = 80 %	0.2288
ถ้า ซื้อหน่อไม้ฝรั่งและแตง แล้ว ซื้อถั่ว	4/14 = 28.6%	4/5 = 80 %	0.2288

5. การรวมกลุ่ม(Clustering)โดยวิธี K-Means

จากที่กล่าวมาแล้วการร่วมกลุ่มเป็นการรวมกลุ่มวัตถุที่มีคุณลักษณะที่คล้ายคลึงกัน ให้รวมอยู่ในกลุ่มเดียวกัน วิธีการที่ใช้ในการรวมกลุ่มมีอยู่หลายวิธีการ เช่น วิธีการรวมกลุ่มแบบมีลำดับชั้น (Hierarchical clustering methods) เหมาะกับการใช้รวมกลุ่มที่วัตถุที่มีจำนวนไม่มากนัก , วิธีการรวมกลุ่มแบบ K-Means เหมาะกับการรวมกลุ่มที่มีวัตถุจำนวนมาก ซึ่งการรวมกลุ่มนั้นไม่อาจที่จะรวมกลุ่มที่เป็นไปได้ทั้งหมด ดังนั้นการรวมกลุ่มควรเลือกวิธีการที่เหมาะสม(Reasonable)ก็เพียงพอ(ปราณี นิลกรณ์ ม.ป.ป : 704) สิ่งที่ใช้ในการรวมข้อมูลคือมาตราที่ใช้วัดความสัมพันธ์หรือมาตราวัดความคล้ายคลึง

ในการวัดความคล้ายคลึงสิ่งสำคัญที่ต้องพิจารณานั้นได้แก่ ชนิดของตัวแปร(ต่อเนื่อง ไม่ต่อเนื่อง) ,มาตราวัด(แบ่งกลุ่ม,อันดับ,อันตรภาค,อัตราส่วน) ,ความรู้เกี่ยวกับข้อมูล (ปราณี นิลกรณ์ ม.ป.ป : 693)เป็นต้น ตัวชี้วัดความคล้ายคลึงที่นิยมใช้คือระยะทางซึ่งมีอยู่หลายรูปแบบเช่น

- ระยะทางยุคลิด (Euclidean distance)

ให้ $X=[x_1,x_2,\cdots,x_n]$ และ $Y=[y_1,y_2,\cdots,y_n]$

$$d_1(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}$$
$$=\sqrt{(x-y)'(x-y)}$$

- ระยะทางยุคลิดกำลังสอง (Square Euclidean distance)

$$d_2(x,y)=(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2$$
$$=(x-y)'(x-y)$$

- ระยะทาง Chebychev

$$d_3(x,y)=Max|x_i-y_i|$$

- ระยะทางมินคอฟสกี(Minkowski metric)

$$d_4(x,y)=\left[\sum_{i=1}^{n}|x_i-y_i|^{\ m}\right]^{\frac{1}{m}}$$

เมื่อ m = 1 $d(x,y)$ วัดระยะทางของบล็อคตึกในเมือง(“city block ” distance) ระหว่าง 2 จุด ถ้า $m=2$ จะเป็นระยะทางยุคลิก ซึ่งถ้า m เปลี่ยนไปจะทำให้น้ำหนักที่ให้แก่ความแตกต่างมากหรือแตกต่างน้อยต่างกันไป(ปราณี นิลกรณ์ ม.ป.ป : 694)

- ระยะทางในเชิงสถิติ

$$d_5(x,y)=\sqrt{(x-y)'A(x-y)}$$

เมื่อ $A=S^{-1}$ โดย S เป็นเมตริกซ์ความแปรปรวนและความแปรปรวนร่วมของตัวอย่าง ซึ่งบางครั้งเราไม่สามารถหาค่า S ได้ จึงนิยมใช้ระยะทางยุคลิดในการวิเคราะห์การรวมกลุ่ม(ปราณี นิลกรณ์ ม.ป.ป : 693)

การรวมกลุ่มแบบ K-Means เหมาะใช้กับกรณีที่มีจำนวนวัตถุที่มากและจะใช้เวลาในการคำนวณที่น้อย โดยจะรวมกลุ่มของวัตถุ K กลุ่ม ซึ่งจำนวน K นั้นอาจจะกำหนดล่วงหน้าหรือกำหนดระหว่างการดำเนินการรวมกลุ่มก็ได้(เช่น กรณีที่เราไม่ทราบจำนวนกลุ่มที่แน่นอน จะทำการวิเคราะห์ด้วย K-Means หลาย ๆ ครั้ง และกำหนดจำนวนกลุ่มที่แตกต่างกันไป เช่น 2 ,3 หรือ 4 แล้วพิจารณากลุ่มที่เหมาะสม หรือ ผู้วิเคราะห์อาจจะใช้วิธีการรวมกลุ่มแบบลำดับชั้นก่อนเพื่อหาว่าควรมีกี่กลุ่มแล้วจึงใช้วิธีแบบ K-Means ในการรวมกลุ่มอีกครั้ง) (กัลยา วานิชย์บัญชา 2546 : 159) วิธีการของ K-Means มีขั้นตอนดังนี้

1. เริ่มจากแบ่งวัตถุที่จะรวมกลุ่มออกเป็น K กลุ่ม

2. นำวัตถุทั้งหมดจัดเข้ากลุ่มที่มีเซนทรอยด์(Centroid)(หรือค่าเฉลี่ย)อยู่ใกล้วัตถุนั้นที่สุด คำนวณเซนทรอยด์ใหม่สำหรับกลุ่มที่มีการเพิ่ม และกลุ่มที่เสียวัตถุไป(ระยะที่ใช้มักจะใช้ระยะทางของยุคลิด ซึ่งจะทำการแปลงค่าสังเกตเป็นค่ามาตราฐาน(Z-score) หรือไม่แปลงก็ได้)

3. ทำในขั้นตอนที่ 2 ซ้ำใหม่จนกระทั่งไม่มีการแปลงกลุ่มของวัตถุหรือครบตามจำนวนรอบที่กำหนด โดยแทนที่จะเริ่มจากการแบ่งวัตถุเป็น K กลุ่มแบบขั้นตอนที่ 1 อาจจะกำหนดเซนทรอยด์เริ่มต้น K กลุ่มและทำในขั้นตอนที่ 2 ก็ได้ การกำหนดกลุ่มให้กับวัตถุในขั้นสุดท้ายจะขึ้นอยู่กับการจัดวัตถุเป็น K กลุ่มตอนต้น และการเลือกเซนทรอยด์ตอนต้นอาจจะเลือกจากประสบการณ์

ตัวอย่างการรวมกลุ่ม

สมมุติมีตัวแปร 2 ตัวคือ x และ y มีวัตถุอยู่ 6 อย่าง ต้องการแบ่งกลุ่มวัตถุเป็น 2 กลุ่ม โดยวิธี K-Means และมีข้อมูลเป็นดังนี้ (Roiger and Geatz 2003 : 84-87)

ตารางที่ 19 ตัวอย่างข้อมูลค่าสังเกตในการทำรวมกลุ่มแบบ K-Means

	ค่าสังเกต	
วัตถุที่	x	y
1	1.0	1.5
2	1.0	4.5
3	2.0	1.5
4	2.0	3.5
5	3.0	2.5
6	5.0	6.0

ที่มา : Richard J. Roiger and Michael W. Geatz , Data Mining : A Tutorial – Based Primer (Minnesota :Pearson Education Inc , 2003) , 84.

เนื่องจากต้องแบ่งกลุ่ม 2 กลุ่ม ซึ่งจะเริ่มจากการแบ่งกลุ่มออกเป็น 2 กลุ่มตามชอบ แล้วหาค่าเฉลี่ยของค่าสังเกตในแต่ละกลุ่ม ซึ่งเป็นพิกัดของเซนทรอยด์ของกลุ่ม โดยให้วัตถุที่ 1 เป็นเซนทรอยด์ของกลุ่มที่ 1 และให้วัตถุที่ 3 เป็นเซนทรอยด์ของกลุ่มที่ 2 ซึ่งจะให้สัญลักษณ์ C_1 และ C_2 แทนกลุ่มที่ 1 และกลุ่มที่ 2 ตามลำดับ C_1 = (1.0,1.5) , C_2 = (2.0,1.5) คำนวณค่าระยะทางยุคลิดของแต่ละวัตถุกับเซนทรอยด์ที่กำหนด

$$d(C_1,1) = 0.00 \qquad d(C_2,1) = 1.00$$
$$d(C_1,2) = 3.00 \qquad d(C_2,2) \approx 3.16$$
$$d(C_1,3) = 1.00 \qquad d(C_2,3) = 0.00$$
$$d(C_1,4) \approx 3.16 \qquad d(C_2,4) = 2.00$$
$$d(C_1,5) \approx 2.24 \qquad d(C_2,5) \approx 1.41$$
$$d(C_1,6) \approx 6.02 \qquad d(C_2,6) \approx 5.41$$

จากระยะทางยุคลิดสามารถกลุ่ม C_1 จะมีวัตถุที่ 1 และ 2

C_2 จะมีวัตถุที่ 3,4,5 และ 6

รอบที่ 2 ทำการคำนวณหาเซนทรอยด์ของ C_1 และ C_2 ใหม่ และทำการหาระยะทางยุคลิดเหมือนในตอนต้น และรวมกลุ่มใหม่ จนกว่าจะไม่มีการเปลี่ยนแปลงของวัตถุในแต่ละกลุ่ม

	พิกัดของเซนทรอยด์	
กลุ่ม	$\bar{x}$	$\bar{y}$
C_1	1.0	3.0
C_2	3.0	3.375

ระยะทางยุคลิดของวัตถุกับกลุ่ม

$d(C_1,1)=1.50$ $d(C_2,1)\approx 2.74$
$d(C_1,2)=1.50$ $d(C_2,2)\approx 2.29$
$d(C_1,3)\approx 1.80$ $d(C_2,3)=2.125$
$d(C_1,4)\approx 1.12$ $d(C_2,4)\approx 1.01$
$d(C_1,5)\approx 2.06$ $d(C_2,5)=0.875$
$d(C_1,6)\approx 5.00$ $d(C_2,6)\approx 3.30$

จากระยะทางยุคลิดสามารถกลุ่ม C_1 จะมีวัตถุที่ 1,2 และ 3

C_2 จะมีวัตถุที่ 4,5 และ 6

รอบที่ 3

	พิกัดของเซนทรอยด์	
กลุ่ม	$\bar{x}$	$\bar{y}$
C_1	1.33	2.50
C_2	3.33	4.00

ระยะทางยุคลิดของวัตถุกับกลุ่ม

$d(C_1,1)\approx 1.05$ $d(C_2,1)\approx 3.42$
$d(C_1,2)\approx 2.03$ $d(C_2,2)\approx 2.38$
$d(C_1,3)\approx 1.20$ $d(C_2,3)\approx 2.83$
$d(C_1,4)\approx 1.20$ $d(C_2,4)\approx 1.42$
$d(C_1,5)=1.67$ $d(C_2,5)\approx 1.53$
$d(C_1,6)\approx 5.07$ $d(C_2,6)\approx 2.60$

จากระยะทางยุคลิดสามารถกลุ่ม C_1 จะมีวัตถุที่ 1,2,3 และ 4

C_2 จะมีวัตถุที่ 5 และ 6

รอบที่ 4

	พิกัดของเซนทรอยด์	
กลุ่ม	$\bar{x}$	$\bar{y}$
C_1	1.50	2.75
C_2	4.00	4.25

ระยะทางยุคลิดของวัตถุกับกลุ่ม

$d(C_1,1) \approx 1.34$ $d(C_2,1) \approx 4.07$

$d(C_1,2) \approx 1.82$ $d(C_2,2) \approx 3.01$

$d(C_1,3) \approx 1.34$ $d(C_2,3) \approx 3.40$

$d(C_1,4) \approx 0.90$ $d(C_2,4) \approx 2.14$

$d(C_1,5) \approx 1.52$ $d(C_2,5) \approx 2.01$

$d(C_1,6) \approx 4.77$ $d(C_2,6) \approx 2.01$

จากระยะทางยุคลิดสามารถกลุ่ม C_1 จะมีวัตถุที่ 1,2,3,4 และ 5

C_2 จะมีวัตถุที่ 6

รอบที่ 5

	พิกัดของเซนทรอยด์	
กลุ่ม	$\bar{x}$	$\bar{y}$
C_1	1.8	2.7
C_2	5	6

ระยะทางยุคลิดของวัตถุกับกลุ่ม

$d(C_1,1) \approx 1.44$ $d(C_2,1) \approx 6.02$

$d(C_1,2) \approx 1.97$ $d(C_2,2) \approx 4.27$

$d(C_1,3) \approx 1.22$ $d(C_2,3) \approx 5.41$

$d(C_1,4) \approx 0.82$ $d(C_2,4) \approx 3.91$

$d(C_1,5) \approx 1.22$ $d(C_2,5) \approx 4.03$

$d(C_1,6) \approx 4.59$ $d(C_2,6) \approx 0$

จากระยะทางยุคลิดสามารถกลุ่ม C_1 จะมีวัตถุที่ 1,2,3,4 และ 5

C_2 จะมีวัตถุที่ 6

ซึ่งไม่มีการเปลี่ยนแปลงกลุ่มของวัตถุแล้วจึงหยุดกระบวนการและได้ของการรวมกลุ่มเป็นดัง C_1 และ C_2

หมายเหตุ มีข้อโต้แย้งเกี่ยวกับการกำหนด K ล่วงหน้าไม่ควรทำ (ปราณี นิลกรณ์ ม.ป.ป : 722)คือ

1. ถ้าจุดเริ่มต้น 2 จุด หรือมากกว่าอยู่ในกลุ่มเดียวกัน ผลของการรวมกลุ่มที่ได้จะจำแนกได้ยากมาก

2. ถ้ามีค่านอกกลุ่ม(outlier) จะทำให้มีอย่างน้อย 1 กลุ่ม ที่วัตถุที่อยู่ในกลุ่มนั้นจะกระจายกันมาก

3. ถึงจะทราบว่าประชากรมี K กลุ่ม แต่การสุ่มตัวอย่างอาจจะทำให้กลุ่มที่เกิดยากไม่ปรากฏในตัวอย่างก็ได้

4. งานวิจัยที่เกี่ยวข้อง

4.1 การรวมกลุ่ม(Clustering)ลักษณะของตลาดการท่องเที่ยวโดยการวิเคราะห์การรวมกลุ่มแบบ K-Means

จากการศึกษาของ Simon Hunson และ Brent Richie (quoted in Daniel T. Larose 2005 :23) ได้ทำความเข้าใจของตลาดการท่องเที่ยวภายในประเทศโดยใช้การวิเคราะห์รวมกลุ่ม (Understanding the domestic market using cluster analysis) โดยตีพิมพ์ในวารสาร Journal of Vacation Marketing เป็นการศึกษาตลาดการท่องเที่ยว โดยอาศัยเทคนิคเหมืองข้อมูล โดยใช้ตัวแบบ CRISP-DM โดยผลการดำเนินการดังนี้

ระยะที่ 1 การทำความเข้าใจธุรกิจหรือการวิจัย

เป็นการศึกษาพฤติกรรมของนักท่องเที่ยวภายใน Alberta ประเทศ แคนนาดา โดยจะสร้างคุณลักษณะของนักท่องเที่ยวภายใน Alberta บนพื้นฐานพฤติกรรมการตัดสินใจของนักท่องเที่ยวเพื่อแบ่งตลาดการท่องเที่ยว โดยจะนำผลการศึกษานี้ช่วยการพัฒนาการส่งเสริมการท่องเที่ยวภายในจังหวัด วัตถุประสงค์หลักเป็นการกำหนดปัจจัยในการเลือกสถานที่เที่ยวภายใน Alberta

ระยะที่ 2 การทำความเข้าใจข้อมูล

ข้อมูลเก็บรวบรวมในปี 1999 โดยใช้วิธีการสำรวจทางโทรศัพท์จากชาว Alberta จำนวน 13,445 คน โดยจะกรองข้อมูลจากผู้ตอบที่อายุเกิน 18 และเที่ยวใช้เวลาว่างในการเที่ยวในระยะ 80 กิโลเมตรสำหรับใช้เวลาการท่องเที่ยวอย่างน้อย 1 คืน โดยข้อมูลที่สมบูรณ์เพียง 3,071 คน จาก 13,445 คนที่จะใช้ในการศึกษาครั้งนี้

ระยะที่ 3 การเตรียมข้อมูล

คำถามที่ใช้ในการสำรวจจะเป็นการแสดงถึงปัจจัยทั้งหมด 13 ปัจจัยที่มีผลต่อการตัดสินใจในการท่องเที่ยวมากที่สุด โดยจะใช้เป็นตัวแปรในการวิเคราะห์รวมกลุ่มจะประกอบไปด้วย คุณภาพของการให้ความสะดวก วันหยุด และสภาพภูมิอากาศ เป็นต้น

ระยะที่ 4 พัฒนาตัวแบบ

Simon Hunson และ Brent Richie เลือกวิธีการรวมกลุ่มโดยเทคนิค K-Means เนื่องจากเป็นเทคนิคที่รวดเร็วและมีประสิทธิภาพซึ่งผู้วิจัยต้องทราบหรือสามารถคาดคะเนจำนวนกลุ่มได้ โดยการศึกษาครั้งนี้จะมีจำนวนกลุ่มอยู่ระหว่าง 2 ถึง 6 กลุ่ม ซึ่งตัวแบบจะกำหนดให้มีจำนวน 5 กลุ่ม เนื่องจากจะสะท้อนความเป็นจริง โดยแต่ละกลุ่มมีลักษณะดังนี้

กลุ่มที่ 1 คือตลาดกลางแจ้งของผู้มีอายุน้อยในเมือง โดยในกลุ่มน่าจะเป็นผู้มีอายุน้อย เพศชายและหญิงเท่ากัน ปฏิทินการศึกษา และงบประมาณ มีผลต่อการตัดสินใจท่องเที่ยว

กลุ่มที่ 2 ตลาดนักท่องเที่ยวที่มีเวลาว่างที่เที่ยวในร่ม จะเป็นกลุ่มที่อายุมากขึ้นมาจากกลุ่มที่ 1 และส่วนมากเป็นผู้หญิง และแต่งงานมีบุตรแล้ว จะมาเที่ยวกับครอบครัวหรือเพื่อน เป็นปัจจัยสำคัญในการวางแผนการเที่ยว

กลุ่มที่ 3 ตลาดที่ให้ความสำคัญของบุตร เป็นคู่สมรสที่แต่งงานมานานและมีบุตรหลายคน สถานที่มีกีฬาสำหรับเด็ก และมีรายการแข่งขันที่มีผลต่อการตัดสินใจต่อการเที่ยวใน Alberta

กลุ่มที่ 4 ตลาดที่ให้ความสำคัญต่อสภาพภูมิอากาศ เพื่อน และราคายุติธรรม เป็นกลุ่มที่มีอายุมากเป็นอันดับที่สอง ส่วนใหญ่จะเป็นกลุ่มของผู้ชาย และเงื่อนไขของสภาพภูมิอากาศมีผลต่อการตัดสินใจในการท่องเที่ยว

กลุ่มที่ 5 ตลาดนักท่องเที่ยวของผู้สูงอายุ ผู้ที่มีอำนาจต่อค่าใช้จ่าย เป็นกลุ่มที่อายุมากที่สุด เป็นผู้พิจารณาต่อค่าใช้จ่ายที่ใช้ และสภาพความปลอดภัย เมื่อตัดสินใจเที่ยวใน Alberta

ระยะที่ 5 การแปลผลและประเมินผล

จะใช้การวิเคราะห์การจำแนก (Discriminant analysis) เพื่อตรวจสอบการสะท้อนความเป็นจริงของการรวมกลุ่ม จากการตรวจสอบการจำแนกพบว่า 93% เป็นการรวมกลุ่มที่ถูกต้อง

ระยะที่ 6 การนำไปใช้

จากการศึกษาพบว่าผลในการดำเนินการส่งเสริมการขายใหม่ที่ชื่อว่า “Alberta ,Made to Order” โดยใช้พื้นฐานในตลาดที่ชนิดของกลุ่มที่ถูกค้นพบในเหมืองข้อมูล มากกว่า 80 โครงการที่ดำเนินการที่ร่วมมือกันระหว่างภาครัฐบาลและภาคธุรกิจ โดย “Alberta ,Made to Order” จะทำการส่งเสริมการตลาดผ่านรายการโทรทัศน์มากกว่า 20 ครั้งโดย 90 % เป็นผู้ที่อายุต่ำกว่า 55 ซึ่งผลทำให้การท่องเที่ยวใน Alberta เพิ่มขึ้นมากกว่า 20 % ที่ชาว Alberta ตัดสินใจเลือกเที่ยวใน Alberta

4.2 การจำแนก(Classification)ลูกค้าที่ไม่สามารถชำระหนี้ได้ของธุรกิจโทรคมนาคม

Daskalaki ,Kopanas , Goudara และ Avouris. (2003 : 239-255) ได้ศึกษาและสร้างระบบสนับสนุนการตัดสินใจในการจัดการลูกค้าที่ไม่สามารถชำระหนี้ได้ สำหรับบริษัทโทรคมนาคมขนาดใหญ่ การพยากรณ์ลูกค้าที่ไม่สามารถชำระหนี้ได้และความแม่นยำในการพยากรณ์ที่ก่อให้เกิดประโยชน์ในการดำเนินธุรกิจ โดยใช้กระบวนการ Knowledge Discovery in Data ซึ่งมีเทคนิคการทำเหมืองข้อมูลเป็นแก่นในการดำเนินงาน ในการศึกษาของ Daskalaki ,Kopanas , Goudara และ Avouris. ได้ใช้ตัวแบบการดำเนินการแบบ 9 ขั้นตอนของ KDD ประโยชน์ที่จะได้รับจากระบบสนับสนุนการตัดสินใจในการจัดการลูกค้าที่ไม่สามารถชำระหนี้ได้ คือ

1. การตรวจพบลูกค้าที่ไม่สามารถชำระหนี้ได้ให้ได้มากที่สุดเท่าที่จะมากได้
2. ลดสัญญาณการเตือนที่ผิดพลาด เช่น จำนวนลูกค้าที่สามารถชำระหนี้ได้ซึ่งถูกจัดให้อยู่ในกลุ่มที่ไม่สามารถชำระหนี้ได้
3. ให้สัญญาณในเวลาที่เหมาะสมถึงผู้ให้บริการในการตั้งเป็นมาตรการป้องกันสำหรับลูกค้าที่มีแนวโน้มไม่สามารถชำระหนี้ได้

ข้อมูลที่ใช้การศึกษาครั้งนี้ได้มาจาก

- ข้อมูลของลูกค้าจากแฟ้มประวัติลูกค้า
- การใช้โทรศัพท์จากศูนย์ข้อมูล
- ข้อมูลการใช้งานสำหรับการเก็บเงินจากระบบสารสนเทศการเก็บเงิน
- รายงานการชำระเงินของลูกค้าจากระบบสารสนเทศการเก็บเงิน
- รายงานการระงับการเชื่อมต่อโทรศัพท์ที่เป็นผลจากการค้างชำระ
- รายการเชื่อมต่อสัญญาณใหม่หลังการชำระเงินแล้ว
- รายงานการยกเลิกสัญญา

เป็นการเก็บข้อมูลจากลูกค้าประมาณ 100,000 ราย ใช้ระยะเวลาในการเก็บข้อมูล 17 เดือน โดยเริ่มจากเดือน 8/1999 จนถึง 2/2001

หลังจากได้ข้อมูลแล้วตัดข้อมูลที่มีมูลค่าค้างชำระที่น้อยกว่า 30 ยูโรทิ้ง นำมากรองข้อมูลและทำความสะอาดข้อมูล(Data cleaning)เพื่อมีความถูกต้อง แล้วใช้สถิติอนุมานมาใช้ เพื่อดูว่าปัจจัยใดสามารถใช้จำแนกลูกค้าที่ไม่สามารถชำระหนี้ได้และสามารถชำระหนี้ได้ออกจากกัน ปัจจัยใดที่ไม่สำคัญ ใช้จำแนกไม่ได้จะตัดทิ้งไป ซึ่งผลจากการดำเนินการทำให้เหลือกรณีที่ศึกษาทั้งหมด

2,066 กรณี ตัวแปรที่ใช้ในการจำแนก 46 ตัวแปร ในการดำเนินงานจะแบ่งข้อมูลเป็น 2 ชุด คือ เซตข้อมูลที่ใช้สร้างตัวแบบ (Training Set) และข้อมูลที่ใช้ในการทดสอบตัวแบบ (Testing Set) เทคนิคเหมืองข้อมูลที่นำมาใช้ในการจำแนกคือ

1. Discriminant analysis แบบ Stepwise โดยใช้วิธี Forward จากการทดสอบ Wilk's Lamda และการทดสอบสถิติ F เพื่อตัดตัวแปรที่มีระดับนัยสำคัญน้อยออก พร้อมทั้งตรวจสอบสหสัมพันธ์ระหว่างตัวแปรอิสระ ตัวแปรใดที่มีความสัมพันธ์กับตัวแปรอื่นที่อยู่ในสมการสูงจะตัดทิ้งไป ผลที่ได้ มี 17 ตัวแปรที่ถูกเลือกมาจาก 46 ตัวแปร

2. Decision tree โดยใช้ตัวแปร 17 ตัวแปรที่ได้จาก Discriminant analysis

3. Neural network แบบ Back propagation โดยจะใช้ตัวแปร 17 ตัวแปรที่ได้จาก Discriminant analysis

ผลจากการวิเคราะห์พบว่า เทคนิค Decision tree ให้ความถูกต้องมากที่สุด และสามารถสรุปประสิทธิ์ภาพในการจำแนกกลุ่ม ได้ดังนี้

ตารางที่ 20 สรุปประสิทธิ์ภาพในการจำแนกกลุ่มของวิธีการต่าง ๆ

ตัววัด	Discriminant analysis(%)	Decision tree(%)	Neural network(%)
สามารถจำแนกลูกค้าที่ไม่สามารถชำระหนี้ได้ถูกต้อง	56.25	59.38	37.50
ความผิดพลาดในการจำแนกกลุ่มลูกค้าที่สามารถชำระหนี้ได้ไปอยู่กลุ่มไม่สามารถชำระหนี้ได้	3.36	1.22	1.68

ที่มา : Daskalaki S. and others ,"Data Mining for Decision Support on Customer Insolvency in Telecommunication Business," European Journal of Operation Research 145 (2003):250.

ในการนำระบบไปใช้งานต้องการที่จะรักษาความสัมพันธ์อันดีกับกลุ่มลูกค้าที่สามารถชำระหนี้ได้นั้น จึงทำการประยุกต์ใช้ทั้ง 3 วิธีการ โดยถ้าลูกค้าที่ถูกจำแนกไปกลุ่มลูกค้าที่ไม่สามารถชำระหนี้ต้องถูกจำแนกทั้ง 3 วิธีการ หรือถ้าลูกค้าที่ถูกจำแนกไปกลุ่มลูกค้าที่สามารถชำระหนี้ต้องถูกจำแนกทั้ง 3 วิธีการ ไม่เช่นจะถูกจำแนกไปยังกลุ่มที่ไม่สามารถจำแนกกลุ่มได้ จากกการดำเนินการ

ดังกล่าวทำให้ความแม่นยำลดลง แต่ก็ทำให้การตัดใจเกี่ยวกับการเตือนที่ผิดพลาด(false alarm)ดีมากขึ้น ได้ทำการประมาณจำนวนเงินที่เป็นหนี้ของกลุ่มลูกค้าไม่สามารถชำระหนี้ได้ทั้งหมด ได้ผลดังนี้

ตารางที่ 21 การประมาณจำนวนเงินที่เป็นหนี้ของกลุ่มลูกค้าไม่สามารถชำระหนี้ได้ทั้งหมด

กลุ่ม(จากการพยากรณ์)	จำนวนเงินที่เป็นหนี้ (%)
ไม่สามารถชำระหนี้ได้(พยากรณ์ถูก)	39
สามารถชำระหนี้ได้(พยากรณ์ผิดคือลูกค้าที่ไม่สามารถชำระหนี้ได้ถูกจำแนกไปอยู่กลุ่มที่สามารถชำระหนี้ได้)	22
ไม่สามารถจำแนกกลุ่มได้	39
รวม	100

ที่มา : Daskalaki S. and others ,"Data Mining for Decision Support on Customer Insolvency in Telecommunication Business," European Journal of Operation Research 145 (2003): 251.

5. บริษัท แฟนซีอารท์ จำกัด

บริษัท แฟนซีอารท์ จำกัด เป็นบริษัทผู้ผลิต นำเข้า และจัดจำหน่ายเสื้อผ้าสำเร็จรูปและของชำร่วย ภายใต้ลิขสิทธ์ของวอทดีสนีย์ นอกจากนั้นยังผลิตสินค้าภายใต้ในนามบริษัทเองด้วย บริษัทตั้งอยู่ที่ 37/4 รัชดาท่าพระ 14 ตลาดพลู กรุงเทพมหานคร และยังมีโรงงานผลิตสินค้าอยู่ที่ อำเภอแม่สอด จังหวัดตาก และ เขต บางขุนเทียน จังหวัด กรุงเทพมหานคร มีบริษัทในเครือคือ บริษัท กิ๊ฟแลนด์ จำกัด (ผลิตของชำร่วย เครื่องเขียน) บริษัท ซันซายน์ จำกัด (โรงงานย้อมผ้า) บริษัท เรนท์โบร์ เอ็มโบเดะริ จำกัด (โรงงานปักลายผ้า) ยังมีบริษัทที่เปิดเป็นร้านค้าอยู่ที่สำเพ็งอีก 4 ร้านค้า บริษัทแฟนซีอาทร์ มีพนักงานทั้งหมดประมาณ 2,000 ราย ในปัจจุบันระบบสารสนเทศที่ใช้ในบริษัท คือโปรแกรม IMEX พัฒนาโดยบริษัท IMEX จำกัด และ โปรแกรม Express พัฒนาโดยบริษัท Indy Soft จำกัด นอกจากนี้ยังมีโปรแกรมที่พัฒนาเองโดยทางบริษัทตามความต้องการของผู้ใช้งาน

บทที่ 3

วิธีดำเนินการพัฒนาระบบ

การทำคลังข้อมูลและเทคนิคการทำเหมืองข้อมูลสำหรับการวิเคราะห์การขาย จะเป็นการพัฒนาระบบคลังข้อมูลและนำข้อมูลที่อยู่ในคลังข้อมูลมาทำการวิเคราะห์ข้อมูลแบบ Multidimensional data analysis โดยการสร้างระบบ OLAP จากนั้นจะนำข้อมูลมาทำเหมืองข้อมูล มีวิธีการดำเนินดังนี้

ขั้นตอนที่ 1

ศึกษาข้อมูลในระบบการขาย เพื่อกำหนดขอบเขตข้อมูลที่จะนำไปใช้หรือที่จะนำเข้าสู่คลังข้อมูล และที่ใช้ในการทำเหมืองข้อมูล

ขั้นตอนที่ 2

วิเคราะห์และออกแบบระบบคลังข้อมูล โดยวิเคราะห์จากความต้องการของผู้ใช้ ปัญหาที่ผู้ใช้พบ และจะนำคลังข้อมูลที่ได้ไปช่วยในการสนับสนุนการตัดสินใจในการบริหารงานที่เกี่ยวกับการขาย

ขั้นตอนที่ 3

นำข้อมูลจากระบบการขาย เข้าสู่ในส่วนของ Data Acquisition และ Data staging area เพื่อการเตรียมพร้อมข้อมูลก่อนนำข้อมูลเข้าคลังข้อมูล โดยในส่วนนี้จะมีการตรวจสอบข้อมูล การทำความสะอาดข้อมูล(Data cleaning) โดยใช้กระบวนการ ETL

ขั้นตอนที่ 4

นำข้อมูลที่ได้เข้าสู่คลังข้อมูลโดยกระบวนการ ETL และออกแบบระบบ OLAP และส่วนติดต่อผู้ใช้ เพื่อให้ผู้ใช้นำข้อมูลในคลังข้อมูลไปใช้ในการวิเคราะห์หรือการใช้ข้อมูลเพื่อการตัดสินใจ

ขั้นตอนที่ 5

นำข้อมูลในคลังข้อมูลไปทำเหมืองข้อมูล โดยใช้ตัวแบบ CRISP-DM(Cross-Industry Standard Process for Data Mining) เพื่อใช้ในการรวมกลุ่มของลูกค้าโดยเทคนิค K-Means สำหรับช่วยในการทำการส่งเสริมการขาย หรือการสร้างความสัมพันธ์กับลูกค้า

เครื่องมือที่ใช้ในการดำเนินการ

1. คลังข้อมูลจะถูกเก็บใน Data Warehouse Database โดยใช้ Microsoft SQL Server 2005 ซึ่งใช้ระบบปฏิบัติการ Microsoft Windows 2003 Server Enterprise Edition

2. ในการทำเหมืองข้อมูลเพื่อการรวมกลุ่มลูกค้า โดยเทคนิค K-Means จะใช้โปรแกรมสำเร็จรูป SPSS

บทที่ 4

ผลการพัฒนาระบบ

4.1 ผลการวิเคราะห์ระบบสารสนเทศการขายในปัจจุบัน

จากการศึกษาในฐานข้อมูลในระบบการขายของบริษัทแฟนซีอารท์พบว่า มีตารางที่ใช้ในการเก็บข้อมูลในระบบการขาย ดังนี้ P01 ประกอบไปด้วย 42 ฟิลด์ , PAY ประกอบไปด้วย 12 ฟิลด์ (ทั้งสองตารางจะเก็บข้อมูลการขาย) , M_CUSTOM ประกอบไปด้วย 31 ฟิลด์(เก็บข้อมูลลูกค้า) , M_STOCK ประกอบไปด้วย 68 ฟิลด์(เก็บข้อมูลสินค้า) ,PGROUP(เก็บข้อมูลกลุ่มลูกค้า) ประกอบไปด้วย 8 ฟิลด์ , SALECODE ประกอบไปด้วย 8 ฟิลด์ (เก็บข้อมูลพนักงานขาย) เมื่อตรวจสอบข้อมูลแล้วจะมีบางฟิลด์ที่ไม่ได้เก็บข้อมูล หรือไม่จำเป็นต่อการเก็บลงคลังข้อมูลหรือไม่จำเป็นต่อการวิเคราะห์ข้อมูล ดังนั้นสามารถสรุปฟิลด์ที่จะใช้ในการเก็บลงคลังข้อมูลในแต่ละตารางได้ดังนี้

M_CUSTOM จะประกอบไปด้วยฟิลด์

ชื่อฟิลด์	คำอธิบาย
CODE	รหัสลูกค้า
NAME	ชื่อลูกค้า
ADDR	ที่อยู่
CR_CTD	วงเงินสินเชื่อ
GRAD	เกรดของลูกค้า

SALECODE จะประกอบไปด้วยฟิลด์

ชื่อฟิลด์	คำอธิบาย
CODE	รหัสพนักงานขาย
NAME	ชื่อพนักงานขาย

PGROUP จะประกอบไปด้วยฟิลด์

ชื่อฟิลด์	คำอธิบาย
CODE	รหัสกลุ่มสินค้า
NAME	ชื่อกลุ่มสินค้า

M_STOCK จะประกอบไปด้วยฟิลด์

ชื่อฟิลด์	คำอธิบาย
CODE	รหัสสินค้า
NAME	ชื่อสินค้า
DETAIL_1	รหัสบาร์โค้ด
PGROUP	รหัสกลุ่มสินค้า

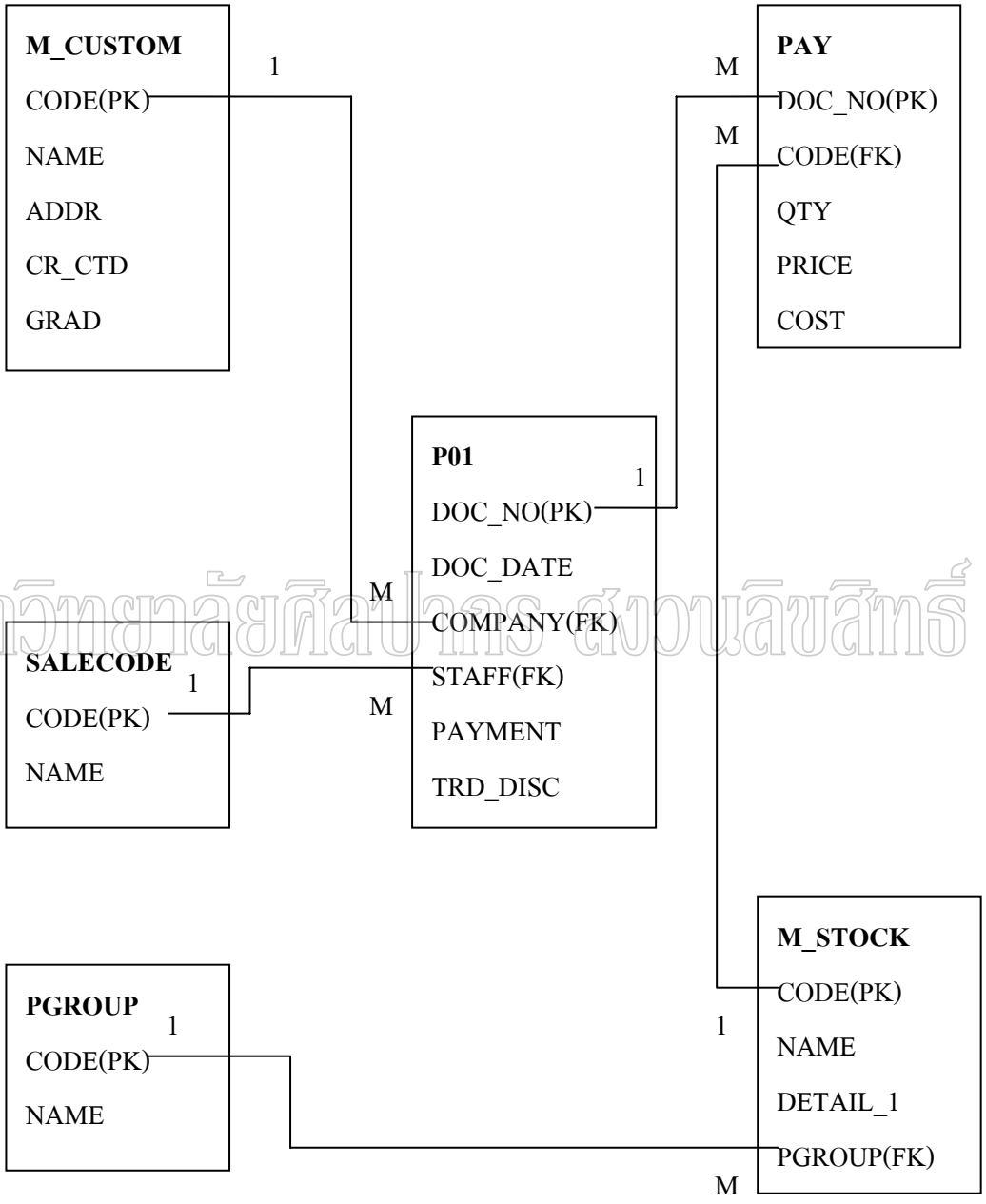
P01 จะประกอบไปด้วยฟิลด์

ชื่อฟิลด์	คำอธิบาย
DOC_NO	เลขที่เอกสาร(ใบกำกับภาษี)
DOC_DATE	วันที่เอกสาร(วันที่ขายสินค้า)
COMPANY	รหัสลูกค้า
STAFF	รหัสพนักงานขาย
PAYMENT	เงื่อนไขการชำระเงิน
TRD_DISC	ส่วนลดการค้า

PAY ประกอบไปด้วยฟิลด์

ชื่อฟิลด์	คำอธิบาย
DOC_NO	เลขที่เอกสาร(ใบกำกับภาษี)
CODE	รหัสสินค้า
QTY	จำนวนที่ขาย
PRICE	ราคาที่ขาย
COST	ต้นทุนสินค้า

จากรายละเอียดข้อมูลของตารางต่าง ๆ สามารถเขียน ER-Diagram ได้ดังนี้



ภาพที่ 18 ER-Diagram ของระบบการขาย

ในการศึกษาครั้งนี้จะแบ่งการแสดงผลการดำเนินการออกเป็น 2 ส่วน คือ ส่วนของการทำคลังข้อมูล และ ส่วนของการทำเหมืองข้อมูล โดยจะใช้ข้อมูลการขายตั้งแต่เดือนตุลาคม พ.ศ. 2547 ถึง เดือนกรกฎาคม พ.ศ. 2548 ในการดำเนินงาน

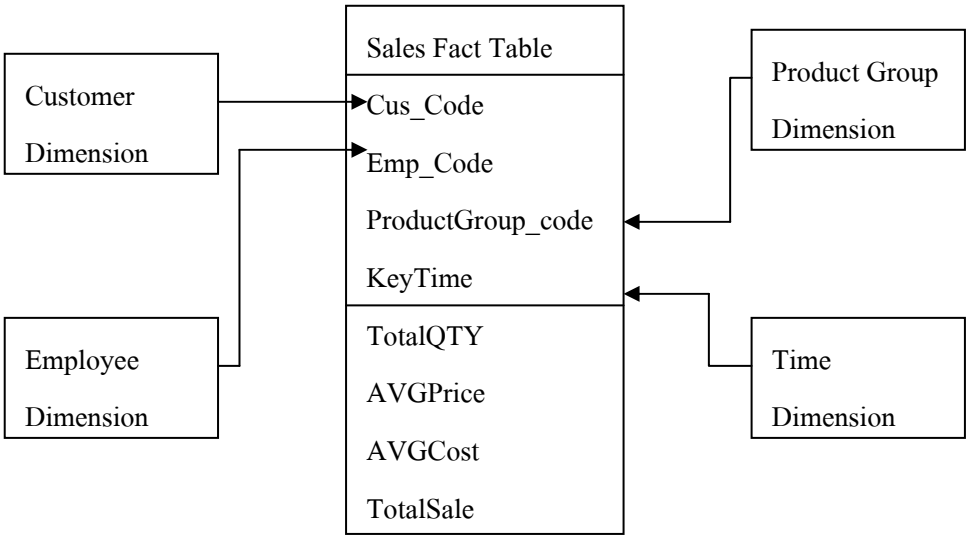
4.2 ส่วนการทำคลังข้อมูล

4.2.1. การออกแบบคลังข้อมูล

จากโครงสร้างของระบบฐานข้อมูลในระบบการขาย สามารถสรุปสารสนเทศที่ต้องการเก็บลงคลังข้อมูลคือ ยอดขาย ปริมาณการขาย ราคาขายโดยเฉลี่ยของสินค้า โดยมองในมุมมองของมิติต่าง ๆ คือ เวลา ลูกค้า กลุ่มของสินค้า และ พนักงานขาย สาเหตุเนื่องจาก

1. ความสมบูรณ์ในการจัดเก็บข้อมูลของระบบการขาย
2. ความถูกต้องของข้อมูล
3. ความต้องการของผู้ใช้ข้อมูล
4. ความง่ายและสะดวกในการจัดเก็บข้อมูลลงในคลังข้อมูล

และจากสารสนเทศดังกล่าวสามารถออกแบบลักษณะของ Schema ของคลังข้อมูลแบบ Star Schema ได้ดังนี้



ภาพที่ 19 Star Schema สำหรับคลังข้อมูลของการขาย

โดยสามารถอธิบายรายละเอียดได้ดังนี้

1. มิติของพนักงานขาย (Employee Dimension : DimEmployee Table)

ชื่อฟิลด์	คำอธิบาย
CODE	รหัสพนักงานขาย
NAME	ชื่อพนักงานขาย

2. มิติของลูกค้า (Customer Dimension : DimCustomer Table)

ชื่อฟิลด์	คำอธิบาย
CODE	รหัสลูกค้า
NAME	ชื่อลูกค้า
ADDR1	ที่อยู่

3. มิติของกลุ่มสินค้า (Product Group Dimension : DimProductGroup Table)

ชื่อฟิลด์	คำอธิบาย
CODE	รหัสกลุ่มสินค้า
NAME	ชื่อกลุ่มสินค้า

4. มิติของเวลา (Time Dimension : DimTime Table)

ชื่อฟิลด์	คำอธิบาย
TimeKey	คีย์เวลา
Doc_Date	วันที่ขาย
DayNumberOfWeek	เลขที่วันภายในสัปดาห์
DayNameOfWeek	ชื่อวันภายในสัปดาห์
MonthNumberOfYear	เลขที่เดือน
MonthNameOfYear	เดือน
NumberOfYear	ปี
QuarterOfYear	ไตรมาส

5. ข้อเท็จจริงของการขาย (Sales Fact Table : FactSale Table)

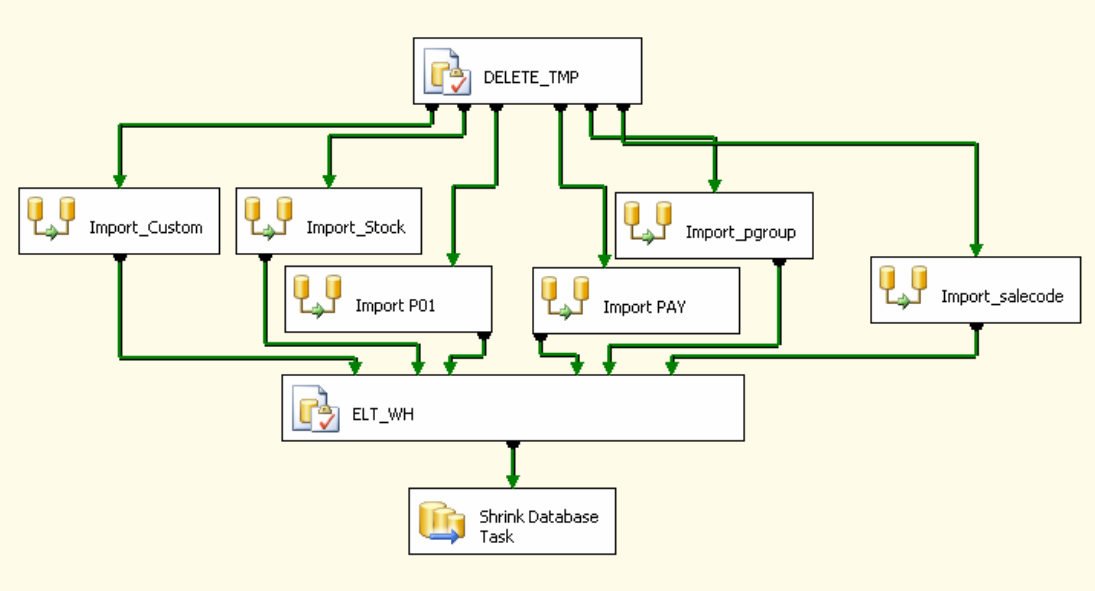
ชื่อฟิลด์	คำอธิบาย
Cus_Code	รหัสลูกค้า
Emp_Code	รหัสพนักงานขาย
ProductGroup_Code	รหัสกลุ่มสินค้า
KeyTime	คีย์เวลา
TotalQTY	จำนวนรวมสินค้าทั้งหมดที่ขายสินค้า (ชิ้น)
AVGPrice	ราคาเฉลี่ยของสินค้า (บาทต่อชิ้น)
AVGCost	ต้นทุนเฉลี่ยของสินค้า (บาทต่อชิ้น)
TotalSale	มูลค่ารวมที่ขายสินค้าทั้งหมด (บาท)

4.2.2 การโอนถ่ายข้อมูล

เป็นนำข้อมูลจากระบบการขาย เข้าสู่ในส่วนของ Data Acquisition และ Data staging area และ การนำข้อมูลจากส่วน Data staging area เข้าสู่คลังข้อมูล ซึ่งในการโอนถ่ายข้อมูลในส่วนแรกจะใช้โปรแกรม Microsft SQL Integration Services (หรือ DTS : Data Tranformation Services) ช่วยในการโอนถ่ายข้อมูลจาก OLTP และในส่วนที่สองจะเป็นการนำข้อมูลจาก Data staging area เข้าสู่คลังข้อมูล ซึ่งจะใช้คำสั่ง SQL (Structure Query Language) ในการดำเนินงาน

ส่วนที่ 1 การนำข้อมูลจาก OLTP เข้าสู่ Data Acquisition และ Data staging area

จากข้อมูลการขายทำให้ทราบว่ามีจำนวนลูกค้าอยู่ในระบบทั้งหมด 3,073 ราย และจำนวนพนักงานขายทั้งหมด 118 ราย มีจำนวนชนิดสินค้าทั้งหมด 89,267 ชนิด จาก 724 กลุ่มสินค้า โดยข้อมูลทั้งหมดเป็นข้อมูลตั้งแต่เริ่มมีการเก็บข้อมูลลงคอมพิวเตอร์ และข้อมูลการขายตั้งแต่เดือนตุลาคม พ.ศ. 2547 ถึง เดือนกรกฎาคม พ.ศ. 2548 จำนวน 96,135 เลขที่เอกสาร และมีจำนวนสินค้าที่ขายทั้งหมด 1,238,185 รายการ ในการนำข้อมูลเข้าต้องการตรวจสอบและแก้ไขข้อมูลที่ผิดพลาด (Missing Value) ในขั้นตอนนี้ด้วยเพื่อเป็นรักษาความถูกต้องของข้อมูลในระดับหนึ่ง โดยมีขั้นตอนการดำเนินงานของ DTS ในโปรแกรมดังนี้



ภาพที่ 20 DTS ที่ใช้ในการโอนถ่ายข้อมูล

โดยการทำงานของ DTS เริ่มจากการลบข้อมูลในที่เก็บข้อมูลชั่วคราว(DELETE_TMP) จะมีคำสั่ง SQL ฝังอยู่ในโมดูล และโมดูล Import_Custom ,Import_Stock ,Import_pgroup ,Import_Salecode,Import_p01,Import_pay จะเป็นการนำข้อมูลเข้าส่วนเก็บข้อมูลชั่วคราว และโมดูล ETL_WH จะเป็นการนำข้อมูลเข้าส่วน Data staging area โดยจะมีการตรวจสอบและแก้ไขข้อมูล ซึ่งจะมีคำสั่ง SQL ฝังอยู่ในโมดูล ในโมดูลสุดท้าย(Shrink Database Task)จะเป็นการกระชับฐานข้อมูลเพื่อให้ฐานข้อมูลมีขนาดเล็กลง

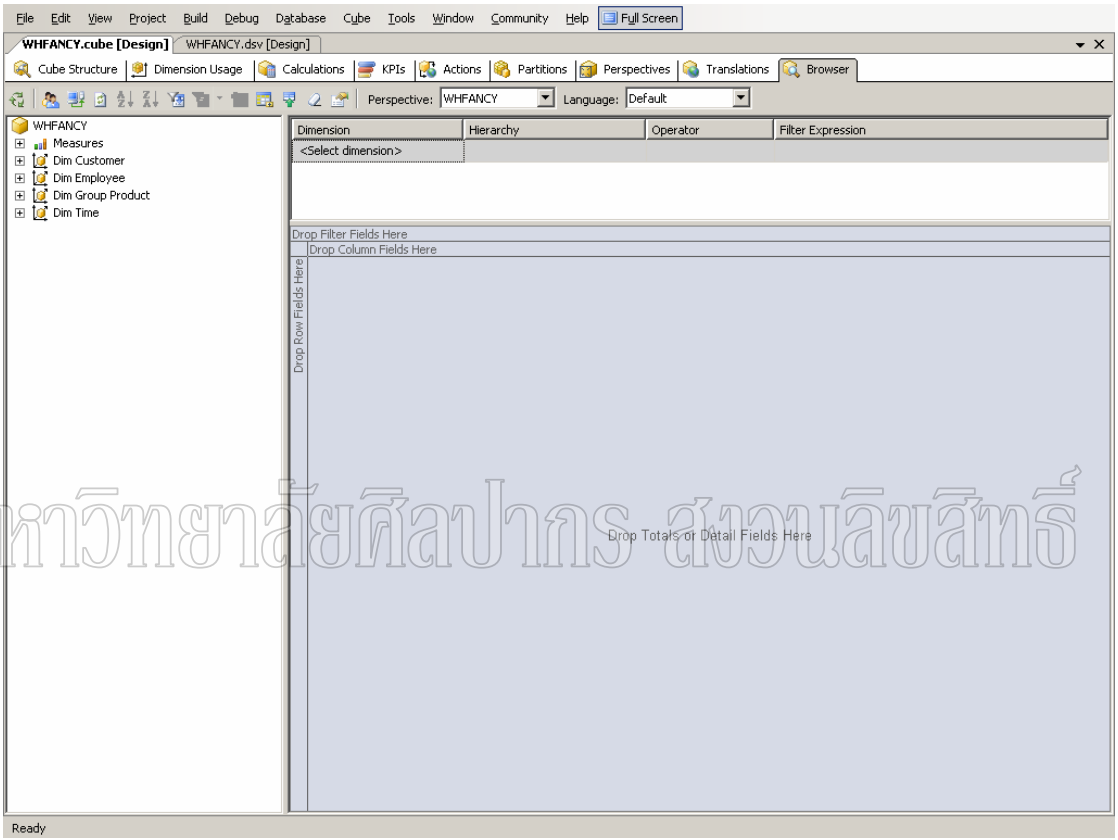
ส่วนที่ 2 การนำข้อมูลจาก Data staging area เข้าสู่ คลังข้อมูล

โดยจะใช้คำสั่ง SQL ในการทำงาน จะมีการแปลงข้อมูลให้มีความสอดคล้องกับโครงสร้างของคลังข้อมูลตาม Schema ที่ออกแบบไว้ พร้อมทั้งตรวจสอบความถูกต้องและแก้ไขข้อมูลที่จะเก็บลงคลังข้อมูลให้มีความถูกต้อง

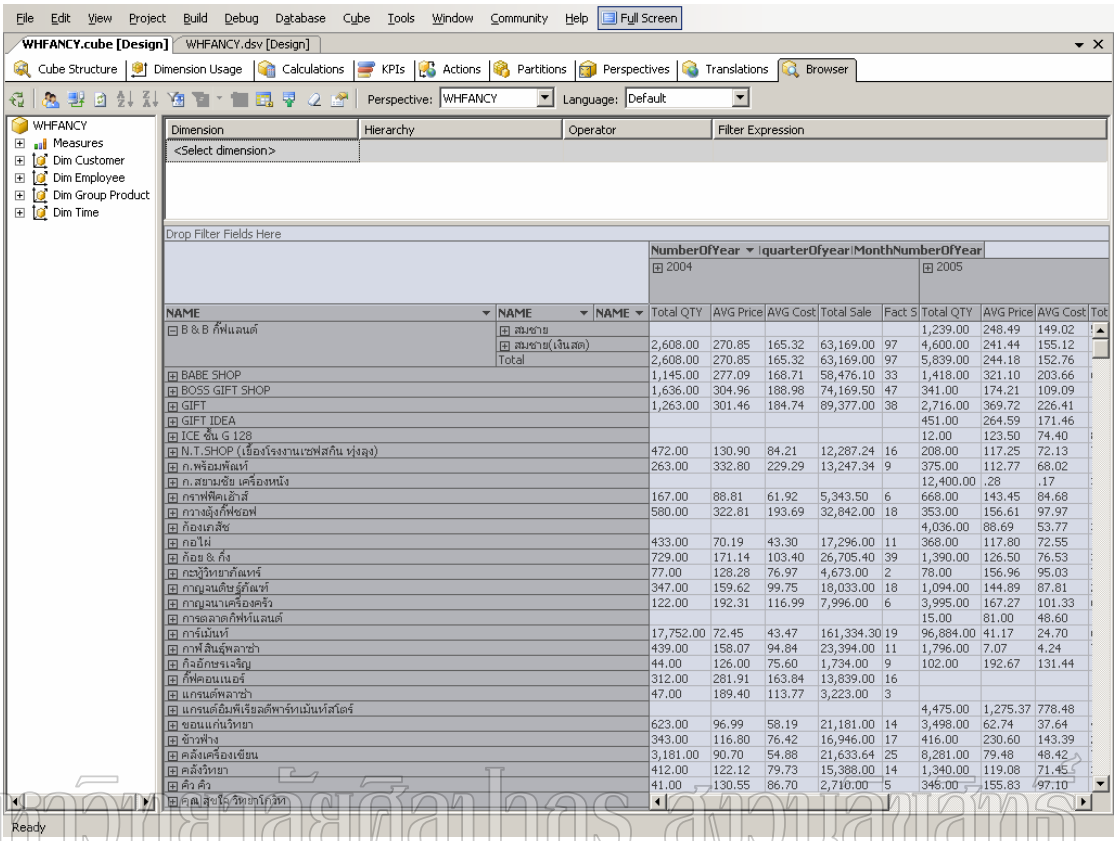
หลังจากนำข้อมูลคลังข้อมูลแล้ว ทำให้ทราบว่ามีข้อเท็จจริง(facts)จำนวน 125,214 ข้อเท็จจริง มีข้อมูลสินค้าที่ยังขายอยู่ 246 กลุ่มสินค้า มีพนักงานขาย 41 ราย และในช่วงเวลาการขายดังกล่าวจะเห็นได้ว่ามีการขายสินค้าให้กับลูกค้าเพียง 760 รายจากจำนวนลูกค้าที่มีอยู่เดิมถึง 3,073 ราย เป็นที่น่าสังเกตว่าเหตุใดลูกค้าเก่าจึงเลิกซื้อสินค้าจากบริษัท ซึ่งในการศึกษาครั้งนี้จะไม่ได้ทำการศึกษากลุ่มลูกค้ากลุ่มนี้

4.2.3. ผลลัพธ์ของคลังข้อมูลโดยการทำ OLAP จากคลังข้อมูล

ในการทำ OLAP นั้นจะใช้โปรแกรม Microsoft Analysis Services ช่วยในการสร้าง Cube และระบบ OLAP มีมิติคือ กลุ่มสินค้า พนักงานขาย ลูกค้า และมิติของเวลา โดยตัววัด (Measure) ของข้อมูลคือ มูลค่าขายรวม จำนวนรวม ราคาเฉลี่ยของสินค้า และ ต้นทุนเฉลี่ยของสินค้า ลักษณะการทำงานของ OLAP เป็นไปดังรูปต่อไปนี้



ภาพที่ 21 การใช้งาน OLAP ก่อนเลือกมิติและตัววัด



ภาพที่ 22 การใช้งาน OLAP หลังเลือกมิติและตัววัด

ในการใช้งานนั้นเราสามารถทำการ Drill down , Roll Up , Slice และการหมุนแกน (Dice) ให้กับ OLAP จะขึ้นอยู่การใช้งานของผู้ใช้งานเพื่อใช้ในวิเคราะห์ข้อมูลสำหรับการตัดสินใจในการบริหารงานต่าง ๆ ข้อมูลที่ได้เป็นข้อมูลเป็นเพื่อวิเคราะห์ในบางครั้งความถูกต้องของข้อมูลจะใช้แทนข้อมูลทางบัญชีไม่ได้ แต่ก็เพียงพอที่จะศึกษาหรือใช้ข้อมูลในวิเคราะห์ได้ และยังได้เปรียบข้อมูลทางบัญชีเนื่องจากข้อมูลที่ได้จะมีความรวดเร็วกว่าข้อมูลกว่าข้อมูลทางบัญชี ทำให้ผู้บริหารสามารถตัดสินใจในการดำเนินนโยบายทางด้านการขายได้ทันต่อเหตุการณ์ อีกทั้งการใช้งานของระบบ OLAP ยังทำให้การประมวลผลการทำงานในการเรียกดูข้อมูลมีความรวดเร็ว

4.3 ส่วนการทำเหมืองข้อมูล

ดังที่ได้กล่าวในบทที่ 3 ในการศึกษาในการทำเหมืองข้อมูล จะใช้ตัวแบบในการศึกษาแบบ CRISP-DM(Cross-Industry Standard Process for Data Mining) โดยมีขั้นตอนในการดำเนินการดังนี้

ระยะที่ 1 การทำความเข้าใจธุรกิจ

ในการขายการสร้างความสัมพันธ์อันดีต่อลูกค้าหรือการส่งเสริมการขายต่าง ๆ ถือเป็นกิจกรรมที่ทำให้สามารถเพิ่มยอดการขาย ทั้งนี้ยังมีวิธีอีกมากมายในการเพิ่มยอดขายสินค้า แต่ถ้าเราทราบลักษณะของกลุ่มลูกค้าจะทำให้การดำเนินกิจกรรมส่งเสริมการขายต่าง ๆ ประสบผลสำเร็จมากขึ้น ถูกกลุ่มเป้าหมายมากขึ้น ในการศึกษาครั้งจึงจะทำการรวมกลุ่มลูกค้าที่มีความคล้ายคลึงไว้ด้วยกัน โดยใช้ข้อมูลการขายของบริษัทในการรวมกลุ่มลูกค้า

ระยะที่ 2 การทำความเข้าใจข้อมูล

ข้อมูลที่ใช้ในการรวมกลุ่มลูกค้าจะใช้ข้อมูลการขายตั้งแต่เดือนตุลาคม พ.ศ. 2547 ถึง เดือนกรกฎาคม พ.ศ. 2548 ในช่วงเวลาดังกล่าวบริษัทขายสินค้าให้กับลูกค้าจำนวน 760 ราย จากจำนวนลูกค้าที่มีทั้งหมด 3,073 ราย มีจำนวนรายการสินค้าที่ขายทั้งหมด 1,238,185 รายการ ดังนั้นในการศึกษาครั้งนี้จะใช้ข้อมูลลูกค้าที่ยังคงซื้อสินค้าจากบริษัทอยู่จำนวนทั้งสิ้น 760 ราย สำหรับกลุ่มลูกค้าที่ไม่มีการซื้อสินค้ากับบริษัทจะไม่นำมาศึกษาในครั้งนี้

ระยะที่ 3 การเตรียมข้อมูล

ตัวแปรที่ใช้ในการวิเคราะห์การรวมกลุ่ม คือตัวแปรที่ได้จากคลังข้อมูลมีดังนี้

1. จำนวนรวมสินค้าทั้งหมดที่ขายสินค้า (TotalQTY)
2. ราคาเฉลี่ยของสินค้า (AVGPrice)
3. ต้นทุนเฉลี่ยของสินค้า (AVGCost)
4. มูลค่ารวมที่ขายสินค้าทั้งหมด (TotalSale)
5. เกรดของลูกค้าที่กำหนด (Grade)
6. วงเงินสินเชื่อในการขายสินค้าของลูกค้า (CR_LTD)

โดยมีข้อมูลรายการสินค้าที่ขายจำนวนทั้งหมด 1,238,185 รายการ และตัวแปรทั้งหมดจะถูกแปลงข้อมูลเป็นค่ามาตรฐาน ก่อนที่จะใช้ในการวิเคราะห์การรวมกลุ่ม โดยสาเหตุในการเลือกตัวแปรในการวิเคราะห์การรวมกลุ่มคือ

1. ข้อมูลที่ใช้ในการวิเคราะห์จะถูกนำจากคลังข้อมูลเป็นหลัก
2. ในการศึกษาครั้งนี้สนใจข้อมูลเชิงปริมาณ ดังนั้นตัวแปรที่เลือกส่วนใหญ่จึงเป็นข้อมูลเชิงปริมาณ
3. ในแต่ละตัวแปรที่เลือกมีความสมบูรณ์และความถูกต้องของข้อมูลสูง

ระยะที่ 4 ตัวแบบ

การศึกษาในการรวมกลุ่มลูกค้า จะใช้เทคนิคแบบ K-Means และใช้ระยะทางแบบยุคลิดในการวิเคราะห์การรวมกลุ่ม เนื่องจากข้อมูลลูกค้ามีจำนวนมาก และการใช้เทคนิคแบบ K-Means เป็นเทคนิคที่มีความรวดเร็วและมีประสิทธิภาพในการวิเคราะห์การรวมกลุ่ม โดยการวิเคราะห์จำเป็นต้องทราบจำนวนกลุ่มหรือสามารถคาดคะเนจำนวนกลุ่มได้ ในการดำเนินการครั้งนี้เราไม่สามารถทราบกลุ่มที่แน่นอนของลูกค้าได้จึงใช้วิธีการวิเคราะห์แบบ K-Means โดยกำหนดให้จำนวนกลุ่มในแต่ละครั้งเป็น 2,3,4,5 และ 6 กลุ่มตามลำดับ ได้ผลตามตารางดังนี้

ตารางที่ 22 จำนวนลูกค้าในแต่ละกลุ่มในการเมื่อกำหนดจำนวนกลุ่มขนาดต่าง ๆ

จำนวนกลุ่ม /จำนวนลูกค้าในแต่ละกลุ่ม	1	2	3	4	5	6
2	3	757				
3	84	673	3			
4	669	45	44	2		
5	48	41	2	2	667	
6	9	26	2	60	2	661

จากข้อมูลในตารางข้างบนเห็นได้ว่าจำนวนกลุ่มที่เหมาะสมในการรวมกลุ่มลูกค้าคือจำนวน 5 กลุ่ม โดยเมื่อรวมกลุ่มลูกค้าให้มีจำนวนกลุ่ม 5 กลุ่ม มีค่าสถิติพื้นฐานได้ดังตารางข้างล่างนี้

ตารางที่ 23 ค่าสถิติพื้นฐานของแต่ละกลุ่มลูกค้า

กลุ่มที่		TotalQTY	AVGPrice	AVGCost	Totalsale	Cr_ltd
1	ค่าเฉลี่ย	34,063.02	181.11	109.76	8,341,633.59	926,829.27
(n = 41 คิดเป็นร้อยละ 5.39 ของจำนวนลูกค้าทั้งหมด)	ส่วนเบี่ยงเบนมาตรฐาน	19,797.50	53.85	33.99	5,644,711.44	543,610.33
2	ค่าเฉลี่ย	1,228,764.50	69.03	41.57	26,791,459.23	0.00
(n = 2 คิดเป็นร้อยละ 0.26 ของจำนวนลูกค้าทั้งหมด)	ส่วนเบี่ยงเบนมาตรฐาน	296,487.75	0.73	0.45	3,254,655.59	0.00

ตารางที่ 23 (ต่อ)

กลุ่มที่		TotalQTY	AVGPrice	AVGCost	Totalsale	Cr_ltd
3	ค่าเฉลี่ย	3,077,031.30	65.44	39.4	52,144,591.32	0
(n = 2 คิดเป็นร้อยละ 0.26 ของจำนวนลูกค้าทั้งหมด)	ส่วนเบี่ยงเบนมาตรฐาน	355,926.30	3.71	2.22	13,659,843.15	0
4	ค่าเฉลี่ย	13,816.73	322.96	191.26	3,898,776.82	57,291.67
(n = 48 คิดเป็นร้อยละ 6.32 ของจำนวนลูกค้าทั้งหมด)	ส่วนเบี่ยงเบนมาตรฐาน	12,332.05	139.92	77.94	3,293,349.66	129,232.42
5	ค่าเฉลี่ย	5,028.43	62.9	38.12	253,880.22	19,070.46
(n = 667 คิดเป็นร้อยละ 87.76 ของจำนวนลูกค้าทั้งหมด)	ส่วนเบี่ยงเบนมาตรฐาน	22,694.18	32.02	19.28	900,852.75	62,260.03
Total(n=760)	ค่าเฉลี่ย	18,454.39	85.73	51.67	1,126,787.46	70,355.26
	ส่วนเบี่ยงเบนมาตรฐาน	171,604.69	82.31	48.37	3,985,514.73	248,963.08

หมายเหตุ ตัวแปร TotalQTY คือ จำนวนรวมสินค้าทั้งหมดที่ขาย (ชิ้น)

ตัวแปร AVGPrice คือ ราคาเฉลี่ยของสินค้า (บาทต่อชิ้น)

ตัวแปร AVGCost คือ ต้นทุนเฉลี่ยของสินค้า (บาทต่อชิ้น)

ตัวแปร TotalSale คือ มูลค่ารวมที่ขายสินค้าทั้งหมด (บาท)

ตัวแปร CR_LTD คือ วงเงินสินเชื่อในการขายสินค้าของลูกค้า(บาท)

ตารางที่ 24 มูลค่ารวมที่ขายสินค้าของแต่ละกลุ่มลูกค้า

กลุ่มที่	จำนวนลูกค้า	มูลค่ารวมที่ขายสินค้าทั้งหมดโดยเฉลี่ยของแต่ละกลุ่ม	มูลค่ารวมที่ขายสินค้าทั้งหมด
1	41	8,341,633.59	342,006,977.19
2	2	26,791,459.23	53,582,918.46
3	2	52,144,591.32	104,289,182.64
4	48	3,898,776.82	187,141,287.36
5	667	253,880.22	169,338,106.74
รวม	760	ค่าเฉลี่ยรวม 1,126,787.46	856,358,472.39

ระยะที่ 5 การแปลผลและประเมินผล

การวิเคราะห์การรวมกลุ่มกำหนดจำนวนกลุ่มเป็น 5 กลุ่ม โดยแต่ละกลุ่มมีรายละเอียดดังนี้

กลุ่มที่ 1

มีลูกค้าที่ถูกรวมเข้ากลุ่มนี้ 41 ราย คิดเป็นร้อยละ 5.39 ของจำนวนลูกค้าทั้งหมด เป็นกลุ่มที่มีค่าเฉลี่ยของวงเงินสินเชื่อในการขายสินค้าของลูกค้ามากที่สุด (ค่าเฉลี่ยเท่ากับ 926,829.27 บาท) จึงน่าจะเป็นกลุ่มลูกค้าที่มีความน่าเชื่อถือมาก โดยลูกค้ากลุ่มนี้จะซื้อสินค้าที่มีราคาเฉลี่ย 181.11 บาท และมีต้นทุนเฉลี่ยของสินค้า 109.76 บาท โดยมีมูลค่ารวมที่ขายสินค้าเท่ากับ 342,006,977.19 บาท เป็นกลุ่มที่ทำรายได้มากที่สุด

กลุ่มที่ 2

มีลูกค้าที่ถูกรวมเข้ากลุ่มนี้ 2 ราย คิดเป็นร้อยละ 0.26 ของจำนวนลูกค้าทั้งหมด เป็นกลุ่มลูกค้าที่เป็นบริษัทในเครือของบริษัทที่เปิดเป็นร้านค้าปลีกซึ่งมีทั้งหมด 4 บริษัท ในกลุ่มลูกค้ากลุ่มนี้มีมูลค่ารวมที่ขายสินค้าเฉลี่ยเท่ากับ 26,791,459.23 บาท โดยลูกค้ากลุ่มนี้จะซื้อสินค้าที่มีราคาเฉลี่ยของสินค้า 69.03 บาท และมีต้นทุนเฉลี่ยของสินค้า 41.57 บาท และในกลุ่มนี้จะไม่มีการให้วงเงินสินเชื่อในการซื้อสินค้า โดยมีมูลค่ารวมที่ขายสินค้าเท่ากับ 53,582,918.46 บาท

กลุ่มที่ 3

มีลูกค้าที่ถูกรวมเข้ากลุ่มนี้ 2 ราย คิดเป็นร้อยละ 0.26 ของจำนวนลูกค้าทั้งหมด เป็นกลุ่มลูกค้าที่เป็นบริษัทในเครือที่เหลืออีก 2 รายของบริษัท แต่เป็นบริษัทที่มียอดขายสูงกว่ากลุ่มบริษัทในเครือในกลุ่มที่ 2 (มีมูลค่ารวมที่ขายสินค้าเฉลี่ยเท่ากับ 52,144,591.32 บาท) โดยลูกค้ากลุ่มนี้จะซื้อสินค้าที่มีราคาเฉลี่ย 65.44 บาท และมีต้นทุนเฉลี่ยของสินค้า 39.40 บาท และในกลุ่มนี้จะไม่มีการให้วงเงินสินเชื่อในการซื้อสินค้า โดยมีมูลค่ารวมที่ขายสินค้าเท่ากับ 104,289,182.64 บาท

กลุ่มที่ 4

มีจำนวนลูกค้าในกลุ่มจำนวน 48 ราย คิดเป็นร้อยละ 6.32 ของจำนวนลูกค้าทั้งหมด เป็นกลุ่มลูกค้าที่จะซื้อสินค้าที่มีราคาสูง(มีราคาเฉลี่ยของสินค้าเท่ากับ 322.96 บาท)และมีต้นทุนเฉลี่ยของสินค้าเท่ากับ 191.26 บาท มีจำนวนรวมของสินค้าโดยเฉลี่ยเท่ากับ 13,816.73 ชิ้น โดยมีมูลค่ารวมที่ขายสินค้าเท่ากับ 187,141,287.36บาท

กลุ่มที่ 5

เป็นกลุ่มที่มีขนาดใหญ่ที่สุดคือมีลูกค้าที่อยู่ในกลุ่มนี้จำนวน 667 ราย คิดเป็นร้อยละ 87.76 ของจำนวนลูกค้าทั้งหมด และเป็นกลุ่มที่มีจำนวนสินค้าที่ขายโดยเฉลี่ย ต้นทุนสินค้าโดยเฉลี่ย ราคาขายเฉลี่ย และมูลค่าขายรวมโดยเฉลี่ยของแต่ละบริษัทน้อยที่สุด(มีจำนวนสินค้าที่ขายเฉลี่ยเท่ากับ 5,028.43 ชิ้น ต้นทุนเฉลี่ยของสินค้าเท่ากับ 38.12 บาท ราคาเฉลี่ยของสินค้าเท่ากับ 62.90 บาท และ

มูลค่ารวมโดยเฉลี่ยเท่ากับ 253,880.22 บาท) แต่อย่างไรก็ตามลูกค้ากลุ่มนี้ทำรายได้ให้กับบริษัทเป็นอันดับ 3 เนื่องจากเป็นกลุ่มขนาดใหญ่

ระยะที่ 6 การนำไปใช้

จากการรวมกลุ่มลูกค้าได้ 5 กลุ่ม จะถูกนำเสนอให้กับผู้บริหารงานด้านการขายเพื่อช่วยในการสนับสนุนการตัดสินใจในการบริหารงานด้านการขาย การกำหนดนโยบายด้านการขาย การบริหารความสัมพันธ์ลูกค้า หรือ การส่งเสริมการขายต่าง ๆ ซึ่งจากการที่เราสามารถรวมกลุ่มลูกค้าช่วยทำให้เราสามารถบริหารงานด้านการขายได้อย่างมีประสิทธิ์ภาพมากขึ้น เช่น การที่จะทำการส่งเสริมการขายเรื่องราคา สามารถกำหนดราคาในการส่งเสริมการขายให้แต่ละกลุ่มแทนที่จะทำการส่งเสริมการขายเหมือนกันหมด ซึ่งจะช่วยให้การส่งเสริมการขายมีประสิทธิ์ภาพมากขึ้น

บทที่ 5

สรุปผลการพัฒนาและข้อเสนอแนะ

สรุปผลการพัฒนาระบบ

ในการศึกษาเรื่องคลังข้อมูลและเทคนิคเหมืองข้อมูลสำหรับการวิเคราะห์ขาย โดยใช้กรณีศึกษาของบริษัทแฟนซีอาร์ทได้แบ่งการดำเนินการออกเป็นสองส่วนคือส่วนของการทำคลังข้อมูลจากระบบการขาย และการทำเหมืองข้อมูลโดยใช้ข้อมูลที่ได้จากคลังข้อมูล

ในการทำคลังข้อมูลจากระบบการขาย เพื่อนำมาช่วยในการวิเคราะห์การขายนั้น พัฒนาในลักษณะของ Star Schema โดยมี Fact Table ที่มีตัววัดคือ จำนวนรวมของสินค้าที่ขาย ราคาสินค้าเฉลี่ย ต้นทุนเฉลี่ยของสินค้า มูลค่ารวมในการขายสินค้า และมีมิติ(Dimension)อยู่ 4 มิติคือ มิติของเวลา มิติของกลุ่มสินค้า มิติของลูกค้า มิติของพนักงานขาย โดยนำเสนอสารสนเทศของการวิเคราะห์การขายอยู่ในรูปของระบบ OLAP (On-Line Analytic Processing) โดยใช้โปรแกรม Microsoft SQL Server 2005 ซึ่งเป็นโปรแกรมจัดการฐานข้อมูลในการดำเนินงาน การใช้งานระบบ OLAP ช่วยให้การวิเคราะห์การขายมีความรวดเร็วในการเรียกดูข้อมูลหรือสารสนเทศ และยังสามารถเรียกดูสารสนเทศในเชิงมิติต่าง ๆที่มีความซับซ้อนได้ เพื่อช่วยสนับสนุนในการตัดสินใจในการบริหารงานด้านการขาย

ส่วนการทำเหมืองข้อมูล ได้ใช้ข้อมูลที่ได้จากคลังข้อมูลซึ่งมีการกรองข้อมูลให้มีถูกต้องของข้อมูลมากขึ้น เนื่องจากเทคนิคในการทำเหมืองข้อมูลนั้นมีมากมายไม่ว่าจะเป็นการใช้เทคนิคทางคณิตศาสตร์ ทางสถิติ หรือทางคอมพิวเตอร์เข้ามาช่วยจัดการข้อมูลที่มีขนาดใหญ่ เพื่อให้ได้สารสนสนเทศหรือสามารถแก้ปัญหาที่ต้องการได้ ในการศึกษาครั้งนี้จึงใช้เทคนิคการรวมกลุ่มแบบ K-Means เพื่อรวมกลุ่มลูกค้าเพื่อช่วยให้สามารถดำเนินนโยบายทางการขาย การทำการส่งเสริมการขาย การสร้างความสัมพันธ์กับลูกค้า หรือ การบริหารงานด้านการขายอื่น ๆ ได้ถูกกลุ่มเป้าหมายมากขึ้น ในการทำเหมืองข้อมูลครั้งนี้ใช้ตัวแบบสำหรับการดำเนินแบบ CRISP-DM(Cross-Industry Standard Process for Data Mining) ซึ่งมีการดำเนินการ 6 ระยะ และในวิเคราะห์การรวมกลุ่มแบบ K-Means ในการศึกษาครั้งนี้ ใช้โปรแกรม SPSS ซึ่งเป็นโปรแกรมวิเคราะห์ข้อมูลทางสถิติในการวิเคราะห์ ผลจากการดำเนินงานสามารถแบ่งกลุ่มลูกค้าได้เป็น 5 กลุ่มลูกค้า จากลูกค้าจำนวน 760 ราย โดยกลุ่มที่ 1 มีลูกค้าที่ถูกรวมเข้ากลุ่มนี้ 41 ราย เป็นกลุ่มที่มีค่าเฉลี่ยของวงเงินสินเชื่อในการขายสินค้ามากที่สุด และเป็นกลุ่มที่ทำรายได้ให้กับบริษัทมากที่สุด กลุ่มที่ 2 และ 3 เป็นกลุ่มลูกค้าที่

เป็นบริษัทในเครือที่เปิดเป็นร้านค้าปลีกซึ่งมีจำนวนสมาชิกในกลุ่มกลุ่มละ 2 ราย แต่มูลค่ารวมที่ขายสินค้าทั้งหมดในกลุ่มที่ 3 จะมากกว่ากลุ่มที่ 2 กลุ่มที่ 4 มีจำนวนลูกค้าในกลุ่มจำนวน 48 ราย เป็นกลุ่มลูกค้าที่จะซื้อสินค้าที่มีราคาเฉลี่ยของสินค้าสูงที่สุด กลุ่มที่ 5 เป็นกลุ่มที่มีขนาดใหญ่ที่สุด คือมีลูกค้าที่อยู่ในกลุ่มนี้จำนวน 667 ราย และเป็นกลุ่มที่มีจำนวนสินค้าเฉลี่ยที่ขายและมูลค่ารวมโดยเฉลี่ยของแต่ละบริษัทน้อยที่สุด

ข้อเสนอแนะเพื่อการวิจัยครั้งต่อไป

1. การทำคลังข้อมูลนั้น การทำความสะอาดข้อมูล(Data Cleaning)หรือการกรองข้อมูลให้มีความถูกต้องนั้นค่อนข้างจะยากลำบากเนื่องมีข้อมูลจากระบบ OLTP(Online Transaction Processing) จำนวนมากและเมื่อนำข้อมูลใหม่เข้าสู่คลังข้อมูลอาจพบข้อผิดพลาดของข้อมูลใหม่ ๆ ดังนั้นจึงควรพัฒนาระบบตรวจสอบข้อมูลแบบอัตโนมัติเพื่อช่วยในการกรองข้อมูล

2. เราสามารถเพิ่มข้อเท็จจริง(Facts) ที่เราสนใจได้ลงในคลังข้อมูล โดยทำการเพิ่มตารางข้อเท็จจริง การเพิ่มตัววัด หรือมิติในการมองข้อมูล ได้ตามต้องการ เพื่อจะได้ข้อเท็จจริงใหม่ ๆ

3. ในการทำเหมืองข้อมูลโดยการรวมกลุ่มลูกค้า ยังไม่ได้มีการทดสอบประสิทธิภาพในการรวมกลุ่ม หรือการสะท้อนความเป็นจริงของกลุ่มลูกค้า เนื่องจากข้อจำกัดเรื่องเวลาในการดำเนินงาน

4. ผู้สนใจในด้านการทำเหมืองข้อมูลอาจจะใช้เทคนิคอื่น ๆ ในการทำเหมืองข้อมูล เช่น การวิเคราะห์สหสัมพันธ์ การวิเคราะห์ถดถอย การทำ Association Rule หรือวิธีอื่น ๆ ในการทำเหมืองข้อมูล เพื่อให้ได้สารสนเทศใหม่ ๆ

5. ในการศึกษาครั้งนี้ไม่ได้ทำการศึกษากลุ่มลูกค้าที่เคยซื้อขายกับบริษัท แต่ในระยะเวลาต่อมาไม่มีการซื้อขายอีก ดังนั้นอาจจะทำการศึกษาข้อมูลลูกค้ากลุ่มนี้ในครั้งต่อไป

บรรณานุกรม

ภาษาไทย

กณิกนันต์ เลียนยี. “การใช้ Robust Regression ในการทำนายปริมาณส่งออกกล้วยไม้ตัดดอก. ” วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาสถิติประยุกต์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร. 2538.

กัลยา วานิชย์บัญชา. การใช้ SPSS for Windows ในการวิเคราะห์ข้อมูล. เล่ม 1 ,การวิเคราะห์สถิติชั้นสูงด้วย SPSS for windows.กรุงเทพฯ :บริษัทธรรมสารจำกัด, 2546 .

กิตติพงศ์ กลมกล่อม. การออกแบบและพัฒนาคลังข้อมูล(Data Warehouse). พิมพ์ครั้งที่ 2 . กรุงเทพฯ : เคทีพี คอมพ์ แอนด์ คอนซัลท์ , 2546 .

ปราณี นิลกรณ์. “เอกสารประกอบการบรรยายวิชาการวิเคราะห์ตัวแปรพหุ.” 2547. (อัดสำเนา)

ไพบูลย์ รัตนประเสริฐ และ วีรานันท์ พงศาภักดี. “เอกสารประกอบการบรรยายวิชาสถิติวิเคราะห์.” 2545. (อัดสำเนา)

วารินทร์ สินสูงสุด . ศิลปะการขาย . กรุงเทพฯ : สายใจ , 2539 .

ศรัณย์ ชูเกียรติ . อีอาร์พีสำหรับธุรกิจขนาดกลางและขนาดย่อมในประเทศไทย-ส่วนงานด้านบัญชีการเงิน . กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย ,2547.

ศิริวรรณ เสรีรัตน์ และคณะ . หลักการตลาด . กรุงเทพ ฯ : บริษัท ธีระฟิล์ม และ ไซเท็กซ์ , 2543 .

สุดาดวง เรืองรุจิระ . หลักการตลาด . พิมพ์ครั้งที่ 9 . กรุงเทพฯ : ประกายพรึก . 2543 .

สุนีย์ พงษ์พินิจภิญโญ. “คลังข้อมูล ” .4 เมษายน 2549.http://www.thaicyberu.go.th/officiatcu/main/2543_09_DatabaseSystem/public_html/lesson15/ms2t1.htm .

ภาษาต่างประเทศ

Berry ,Michael J.A. ,and Gordon S. Linoff . Data Mining Techniques :for Marketing, Sales, and Customer Relationship Management. 2nd edition . Indiana : Wiley Publishing Inc, 2004 .

Daskalaki, Sofia ,and others. “Data Mining for Decision Support on Customer Insolvency in Telecommunication Business.” European Journal of Operation Research 145(2003):239-255.

Larose ,Daniel T. Discovering Knowledge in Data :an Introduction to Data Mining .New Jersey : John Wiley & Sons Inc, 2005.

Roiger ,Richard J. ,and Michael W. Geatz . Data Mining : A Tutorial – Based Primer. Minnesota : Pearson Education Inc, 2003.

Rygielski, Chris, Jyun-Cheng Wang ,and David C. Yen. "Data Mining Techniques for Customer Relationship Management. " Technology In Society 24 (2002) : 483-502 .

Hudson, Simon ,and Brent Richie. "Understanding the Domestic Market Using Cluster Analysis a Case Study of the Marketing Efforts of Travel Alberta." Journal of Vacation Marketing ,no. 3 (2002) :263-276.

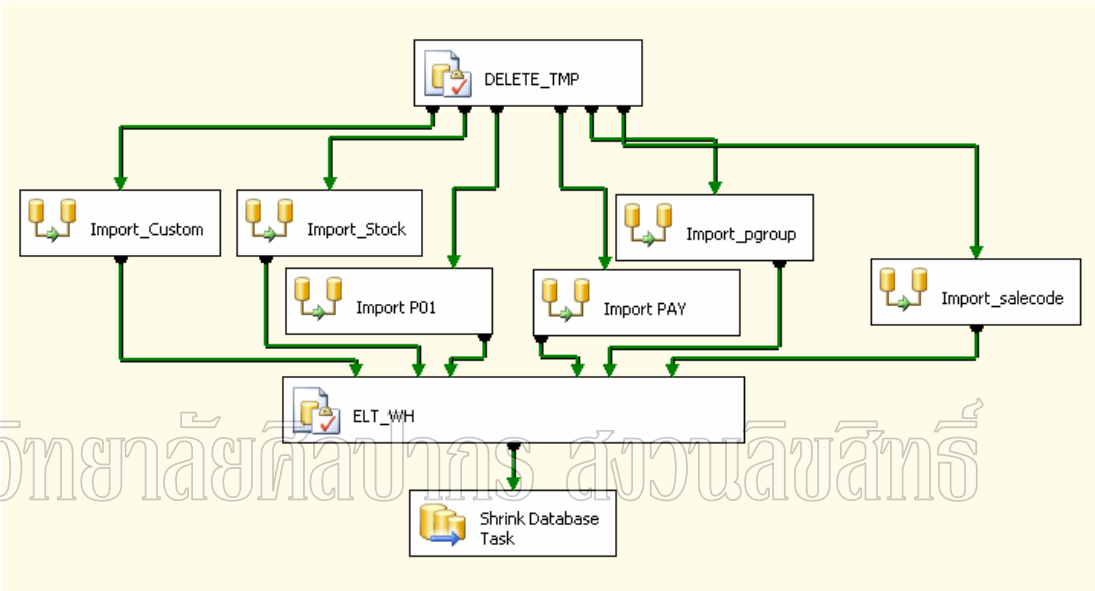
Witten ,Ian H. ,and Eibe Frank . Data Mining :Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco : Elsevier Inc, 2005.

Youness ,Sakhr . Professional Data Warehousing with SQL Server 7.0 and OLAP Services . Birmingham : Wrox Press , 2005.

ภาคผนวก

การโอนถ่ายข้อมูล

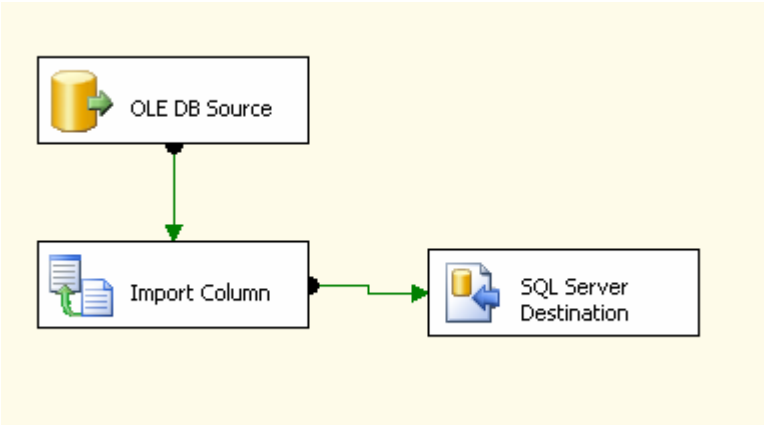
โมดูลการทำงานของการโอนถ่ายข้อมูลจากฐานข้อมูลปฏิบัติงาน (Operational database) เข้าสู่ฐานข้อมูลคลังข้อมูล(Data warehouse database)



คำสั่งในโมดูล DELETE_TMP ซึ่งลบข้อมูลออกจากที่เก็บข้อมูลชั่วคราว

```
Use WHFANCY
Delete From Custom_tmp
Delete From Stock_tmp
Delete From PGROUP_TMP
Delete From SaleCode_TMP
Delete From P01_TMP
Delete From Pay_TMP
GO
```

โมดูล Import_Custom ,Import_Stock ,Import_pgroup ,Import_Salecode,Import_p01,Import_pay จะเป็นการนำข้อมูลจากฐานข้อมูลปฏิบัติงานเข้าสู่ที่เก็บข้อมูลชั่วคราว โดยในแต่ละโมดูลจะมีลักษณะการทำงานที่เหมือนกันคือ



คำสั่งใน โมดูล ELT_WH เป็นการนำข้อมูลเข้าสู่ส่วน Data staging area

```
Use WHFANCY
-- Delete Data Of Master Table From Data Warehouse
Delete From WH_CUSTOM
Delete From WH_STOCK
Delete From WH_PGROUP
Delete From WH_SALECODE
-- Check PAY and P01
delete from WH_P01
        where Doc_no In(Select DISTINCT Doc_no From P01_Tmp)
delete from WH_PAY
        where Doc_no in (select DISTINCT Doc_no From Pay_TMP)
Go
---- Tranformation Data
Insert Into WH_CUSTOM (CODE,NAME,ADDR1,GRADE,CR_LTD)
        Select DISTINCT CODE,NAME,ADDR1,'aGrade'=
        case
                when Grade = 'A' Then 2
                when Grade = 'B' Then 1
                else 0
        end,CR_LTD
        From Custom_tmp
```

```
Insert Into WH_P01(DOC_NO,DOC_Date,COmpany,Staff,Payment,Trd_DISC)
        (select DISTINCT  Doc_no ,DOc_Date,Company,Staff,'aPayment' =
        case
                when substring(Payment,1,2) > '00' And substring(Payment,1,2) < '99' then
cast(substring(Payment,1,2) as int)
                else 0
        end
,'aTRD_DISC' =
        case
                when trd_disc is null then 0.00
                when len(rtrim(ltrim(trd_disc))) = 3 then cast(substring(trd_disc,1,2) as real)
                when len(rtrim(ltrim(trd_disc))) = 2 then cast(substring(trd_disc,1,1) as real)
                when len(rtrim(ltrim(trd_disc))) = 6 then cast(substring(trd_disc,1,2) as real) + (
cast(substring(trd_disc,1,2) as real)  * cast(substring(trd_disc,5,1) as real) /100)
                else 0.00
        end
from p01_tmp  where doc_Date is not null)
Insert Into WH_STOCK(Code,Name,Detail_1,PGROUP)
        Select DISTINCT Code,Name,Detail_1,PGROUP from Stock_tmp
Insert Into WH_Salecode(Code,Name)
        select DISTINCT Code,Name From Salecode_Tmp
Insert Into WH_Pay(DOC_NO,Code,Cost,Qty,Price)
        (Select DISTINCT Doc_No,Code,'Cost' = Case
                when cost < price and cost <> 0  then Pay_tmp.Cost
                when cost = 0 then Pay_tmp.Price - (Pay_tmp.Price * 0.4)
                when cost >= price then Pay_tmp.Price - (Pay_tmp.Price * 0.4)
        end
        ,Qty,Price From Pay_tmp where (Qty * Price) <> 0)
Insert Into WH_PGROUP(Code,Name)
        Select DISTINCT Code,Name From PGROUP_TMP
```

คำสั่ง ในการนำข้อมูลจาก Data staging area เข้าสู่ฐานข้อมูลคลังข้อมูล

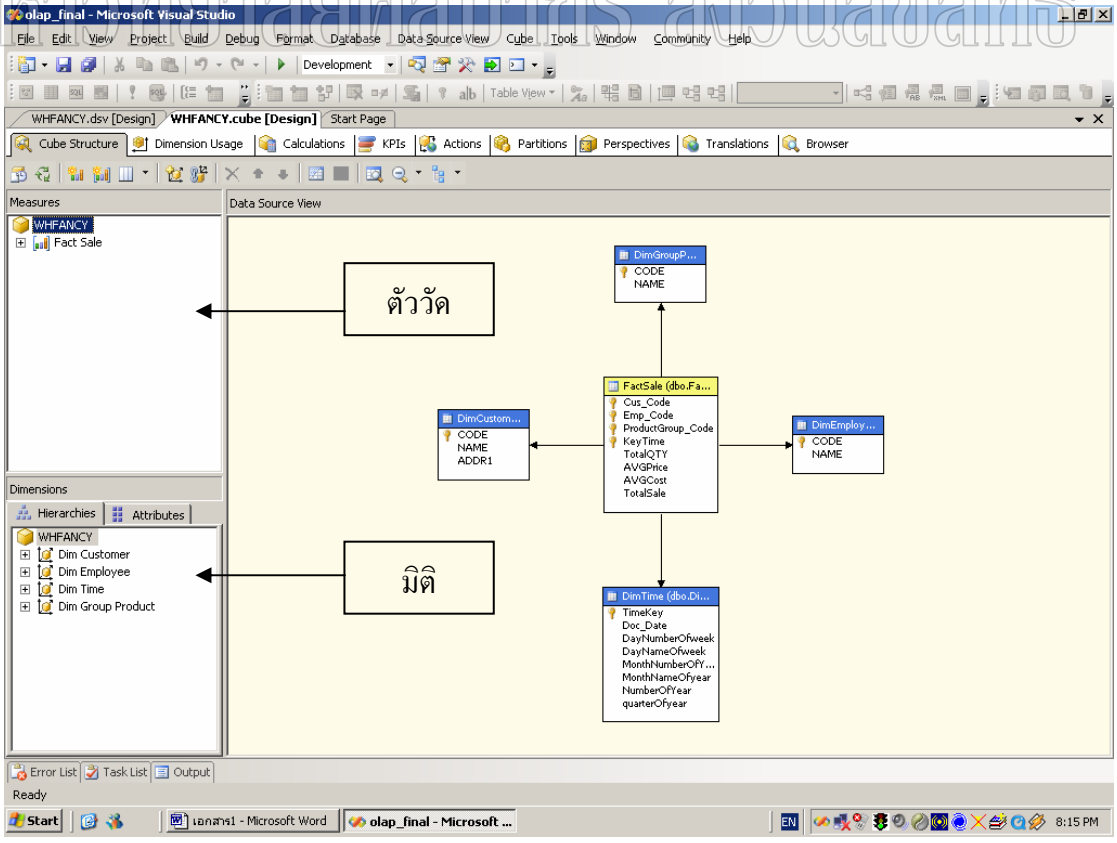
```
Use WHFancy
-- Delete Dimension
Delete from DimCustomer
Delete From DimEmployee
Delete from DimGroupProduct
Delete From Date_TMP
Insert Into Date_tmp(DOC_DATE)
        Select Distinct Doc_Date From WH_P01 Group By doc_Date
Go
-- Insery Data To Dimension
Insert Into DimCustomer(Code,Name,Addr1)
        Select DISTINCT Code,Name,Addr1 From WH_Custom
Insert Into DimEmployee(Code,Name)
        Select DISTINCT Code,Name From WH_SaleCode
Insert Into DimGroupProduct(Code,Name)
        Select DISTINCT Code,Name From WH_PGROUP
go
Insert Into
DimTime(TimeKey,Doc_Date,DayNumberOfWeek,DayNameOfWeek,MonthNumberOfYear,Mo
nthNameOfYear,NumberOfYear,QuarterOfYear)
        select DISTINCT 'TimeKey' = ROW_NUMBER()over (order by doc_date),doc_date,
                'DayNumberOfWeek'= datepart(dw,Doc_date),
                'DayNameOfweek'= case
                        when  datepart(dw,Doc_date) = 1 then 'วันอาทิตย์'
                        when  datepart(dw,Doc_date) = 2 then 'วันจันทร์'
                        when  datepart(dw,Doc_date) = 3 then 'วันอังคาร'
                        when  datepart(dw,Doc_date) = 4 then 'วันพุธ'
                        when  datepart(dw,Doc_date) = 5 then 'วันพฤหัสบดี'
                        when  datepart(dw,Doc_date) = 6 then 'วันศุกร์'
                        when  datepart(dw,Doc_date) = 7 then 'วันเสาร์'
```

```
                else 'Not Day'
            end ,'MonthNumberOfYear' = datepart(month,doc_date),
            'MonthNameOfYear' = case
                when datepart(month,doc_date)= 1 then 'มกราคม'
                when datepart(month,doc_date)= 2 then 'กุมภาพันธ์'
                when datepart(month,doc_date)= 3 then 'มีนาคม'
                when datepart(month,doc_date)= 4 then 'เมษายน'
                when datepart(month,doc_date)= 5 then 'พฤษภาคม'
                when datepart(month,doc_date)= 6 then 'มิถุนายน'
                when datepart(month,doc_date)= 7 then 'กรกฎาคม'
                when datepart(month,doc_date)= 8 then 'สิงหาคม'
                when datepart(month,doc_date)= 9 then 'กันยายน'
                when datepart(month,doc_date)= 10 then 'ตุลาคม'
                when datepart(month,doc_date)= 11 then 'พฤศจิกายน'
                when datepart(month,doc_date)= 12 then 'ธันวาคม'
                else 'Not Mounth'
            end,'NumberOfYear' = datepart(year,doc_date),
            'QuaterOfyear'=datepart(quarter,Doc_date)
        from Date_TMP
        where doc_date not in (Select Doc_Date From DimTime)
go
--Insert FactSale [Sale Fact Table]
Insert Into
FactSale(Cus_Code,Emp_Code,ProductGroup_Code,Keytime,TotalQty,AVGPrice,TotalSale,AV
GCost)
        (SELECT  DISTINCT    WH_CUSTOM.CODE AS Cus_Code,
WH_SALECODE.CODE AS Emp_Code, WH_PGROUP.CODE AS ProductGroup_Code,
DimTime.TimeKey As TimeKey ,
            SUM(WH_PAY.QTY) AS TotalQty, AVG(WH_PAY.PRICE) AS AVGPrice,
SUM(WH_PAY.QTY * WH_PAY.PRICE) AS 'TotalSale', AVG(WH_PAY.COST)
            AS AVGCost
```

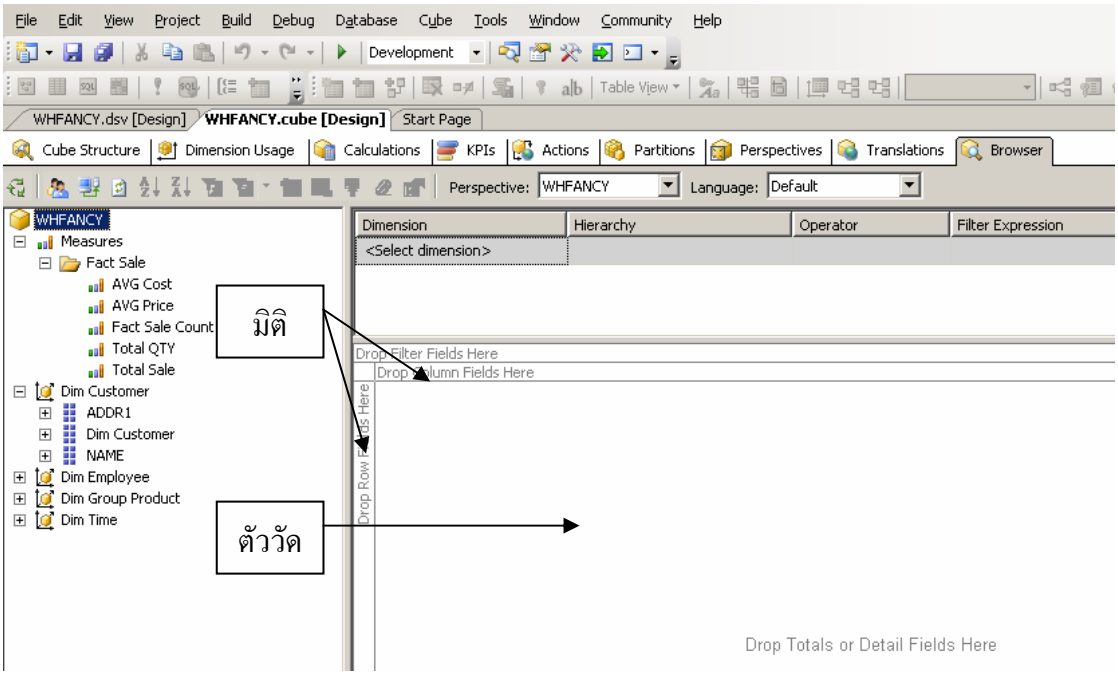
```
FROM        WH_PAY INNER JOIN
            WH_P01 ON WH_PAY.DOC_NO = WH_P01.DOC_NO INNER JOIN
            WH_CUSTOM ON WH_P01.COMPANY = WH_CUSTOM.CODE INNER JOIN
            WH_SALECODE ON WH_P01.STAFF = WH_SALECODE.CODE INNER JOIN
            WH_STOCK ON WH_PAY.CODE = WH_STOCK.CODE INNER JOIN
            WH_PGROUP ON WH_STOCK.PGROUP = WH_PGROUP.CODE INNER JOIN
            DimTime ON WH_P01.DOC_DATE = DimTime.Doc_Date
      GROUP BY WH_CUSTOM.CODE, WH_SALECODE.CODE, WH_PGROUP.CODE,
DimTime.TimeKey
      Having SUM(WH_PAY.QTY * WH_PAY.PRICE) <> 0 )
Go
```

การใช้งาน OLAP

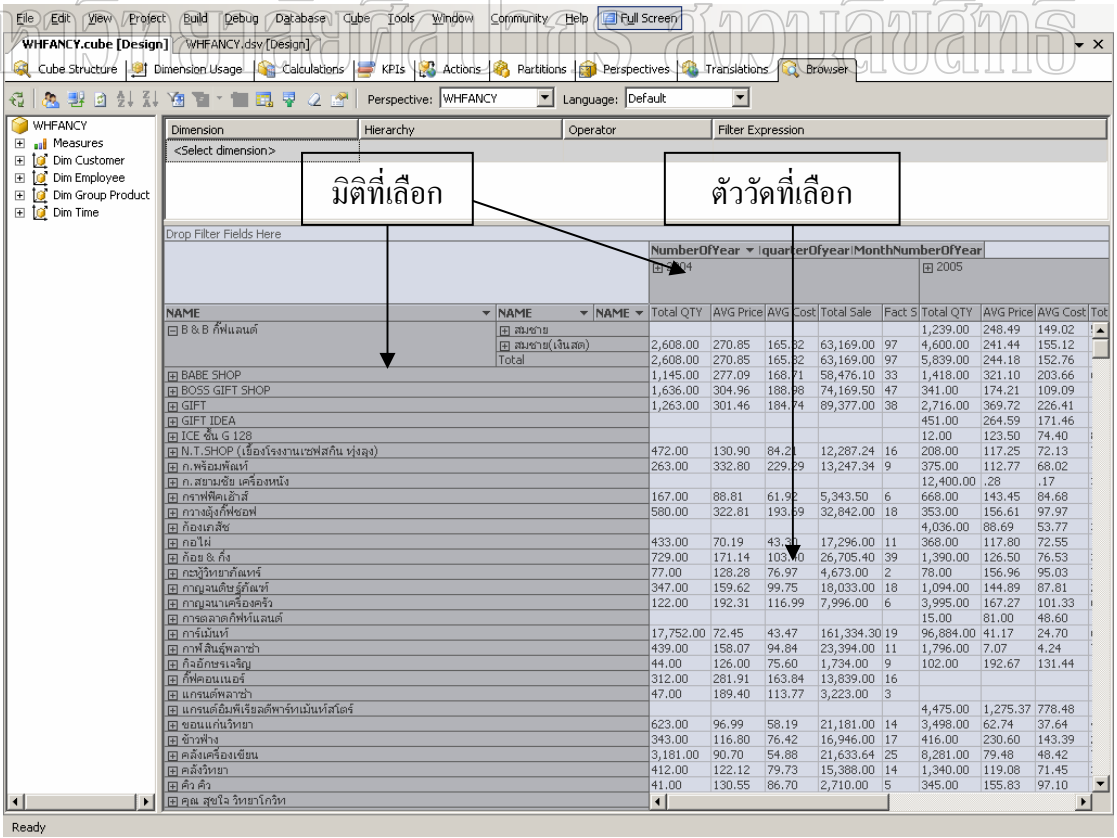
กำหนด Schema ในระบบ OLAP



การเลือกข้อมูล



การใช้งานหลังเลือกข้อมูลและประมวลผล



ประวัติผู้วิจัย

ชื่อ - สกุล นาย บวร น้อยแสง

ที่อยู่ 1 หมู่ที่ 4 แขวงหนองค้างพลู เขตหนองแขม จังหวัด กรุงเทพมหานคร 10160

ประวัติการศึกษา

พ.ศ. 2544 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขา สถิติประยุกต์ จาก สถาบันราชภัฏธนบุรี เขต ธนบุรี กรุงเทพ ฯ

พ.ศ. 2545 ศึกษาต่อระดับปริญญาวิทยาศาสตรมหาบัณฑิต สาขาคณิตศาสตร์และเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร พระราชวังสนามจันทร์ นครปฐม