

พิมพ์ครั้งที่

2

คลังข้อมูล เพื่อการจัดการข้อมูลสมัยใหม่

DATA WAREHOUSING
FOR MODERN DATA
MANAGEMENT

รัฐสฤทธิ สุขะหุต

คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

คำนำ (พิมพ์ครั้งที่ 1)

ปัจจุบันในยุคสมัยของโลกสังคมดิจิทัลและสื่อออนไลน์ที่มีการพึ่งพาเทคโนโลยีสารสนเทศมากยิ่งขึ้น วิทยาการข้อมูลได้เข้ามามีบทบาทสำคัญองค์กรต่างๆ ทั้งหน่วยงานในภาครัฐและเอกชนที่มีการจัดเก็บข้อมูลในรูปแบบของสื่ออิเล็กทรอนิกส์ต่างให้ความสำคัญต่อการจัดเก็บข้อมูลสำหรับนำข้อมูลมาใช้เพื่อประกอบการวางแผน วิเคราะห์ และเป็นข้อมูลสนับสนุนการตัดสินใจของผู้บริหาร แนวคิดของการจัดการข้อมูลสมัยใหม่เป็นศาสตร์ที่ว่าด้วยการบริหารจัดการข้อมูลอย่างเป็นระบบ เป็นการผสมผสานระหว่างวิทยาการคอมพิวเตอร์ หลักสถิติ คณิตศาสตร์ และซอฟต์แวร์ และต้องอาศัยความรู้ทางด้านการจัดการแบบบูรณาการ ซึ่งจำเป็นต้องให้ความสำคัญต่อการออกแบบโครงสร้างและสถาปัตยกรรมข้อมูลที่ดี เพื่อให้การใช้ข้อมูลสารสนเทศเกิดประโยชน์สูงสุดต่อองค์กร

หนังสือคลังข้อมูลเพื่อการจัดการข้อมูลสมัยใหม่เล่มนี้ได้เรียบเรียงจากเอกสาร ตำรา และการค้นคว้าความรู้เพิ่มเติมจากแหล่งข้อมูลต่างๆ รวมถึงการสอดแทรกความรู้และประสบการณ์ของผู้เขียน เพื่อให้เนื้อหามีความทันสมัยโดยถ่ายทอดเป็นหนังสือเล่มนี้ ซึ่งมุ่งเน้นการนำเสนอแนวคิดของการพัฒนาคลังข้อมูล เหตุผลและความจำเป็นของการพัฒนาคลังข้อมูลเพื่อมุ่งเน้นการพัฒนาข้อมูลไปสู่ข้อมูลสารสนเทศสำหรับผู้บริหาร และได้อธิบายถึงขั้นตอนการออกแบบคลังข้อมูลโดยอาศัยไคเมนชันโมเดล ตลอดจนแนวทางการประยุกต์ใช้ข้อมูลที่จัดเก็บภายในคลังข้อมูล การประมวลผลข้อมูลเชิงมิติโดยอาศัยเครื่องมือโอแลป นอกจากนี้ยังได้อธิบายถึงภาษาเอ็มดีเอ็ชซึ่งเป็นโปรแกรมภาษาที่ใช้ในการค้นคืนข้อมูลเชิงมิติเพื่อนำไปสู่การใช้ประโยชน์จากข้อมูลเชิงสารสนเทศโดยใช้เครื่องมือธุรกิจอัจฉริยะ และท้ายสุดได้อธิบายถึงการบำรุงดูแลรักษาระบบคลังข้อมูล การบริหารความเสี่ยงและความต่อเนื่องทางธุรกิจ ซึ่งต้องคำนึงถึงความมั่นคงปลอดภัยของข้อมูล เนื้อหาของหนังสือเล่มนี้ได้มีการเรียบเรียงและนำเสนออย่างละเอียดทุกขั้นตอนพร้อมภาพประกอบเนื้อหาแบ่งเป็นบทต่างๆ สุดท้ายนี้ผู้เขียนหวังว่าหนังสือเล่มนี้จะเป็นประโยชน์สำหรับอาจารย์ นักศึกษาสาขาวิชาวิทยาการคอมพิวเตอร์ วิศวกรรมคอมพิวเตอร์ และสาขาเทคโนโลยีสารสนเทศ รวมถึงผู้ที่สนใจในเนื้อหาที่เกี่ยวกับการคลังข้อมูลและเทคโนโลยีฐานข้อมูลต่อไป

ผู้ช่วยศาสตราจารย์ ดร.รัฐสิทธิ์ สุขะหุด

คำนำ (พิมพ์ครั้งที่ 2)

ในโลกของเทคโนโลยีข่าวสารที่ข้อมูลมีบทบาทสำคัญต่อการดำเนินธุรกิจ ทำให้หลายองค์กรปัจจุบันเริ่มหันมาให้ความสนใจต่อการจัดเก็บข้อมูลของตนเองมากขึ้น การให้ความสำคัญต่อการบริหารจัดการข้อมูลเหล่านี้จึงเป็นสิ่งจำเป็น ข้อมูลที่เก็บสะสมมาตั้งแต่อดีตกลายมาเป็นข้อมูลขนาดใหญ่ที่มีปริมาณมหาศาล ข้อมูลเหล่านี้กลายมาเป็นสินทรัพย์ที่สำคัญขององค์กร หากนำมาใช้ให้เกิดประโยชน์ ข้อมูลสารสนเทศที่ผ่านกระบวนการสังเคราะห์ทำให้ผู้บริหารขององค์กรเข้าใจสถานการณ์และผลประกอบการขององค์กรตนเอง สามารถนำเอาข้อมูลมาใช้ในการวางแผน การวิเคราะห์ และตัดสินใจ รวมถึงการใช้ประโยชน์ข้อมูลเพื่อสร้างความได้เปรียบต่อคู่แข่ง

คลังข้อมูลเป็นแนวคิดของการบริหารจัดการข้อมูลขนาดใหญ่เพื่อการนำข้อมูลไปใช้ประโยชน์เพื่อการวางแผนและสนับสนุนการตัดสินใจสร้างความได้เปรียบเชิงกลยุทธ์ในการแข่งขันในโลกธุรกิจ คลังข้อมูลเป็นการออกแบบโครงสร้างข้อมูลที่รองรับข้อมูลที่มีปริมาณมาก เป็นข้อมูลที่มีการเชื่อมโยงจากข้อมูลหลายแหล่ง สามารถเพิ่มประสิทธิภาพในการวิเคราะห์ การวิเคราะห์ คลังข้อมูลเป็นการจัดเก็บข้อมูลย้อนหลังในอดีต ต่างจากฐานข้อมูลระดับปฏิบัติการที่ข้อมูลมีความเป็นปัจจุบัน เนื้อหาของหนังสือเล่มนี้ได้อธิบายภาพรวมของความสำคัญของการบริหารจัดการข้อมูลขนาดใหญ่ ประกอบด้วยขั้นตอนการวิเคราะห์ ออกแบบ และพัฒนาไปสู่โมเดลการสร้างมุมมองข้อมูลเชิงมิติ ตลอดจนการนำเอาข้อมูลที่จัดเก็บภายในคลังข้อมูลมาใช้ประโยชน์ ทั้งการวิเคราะห์ข้อมูลขนาดใหญ่โดยใช้ภาษาเอ็มทีเอ็กซ์ และการพัฒนาไปสู่ข้อมูลสารสนเทศสำหรับผู้บริหารโดยใช้เครื่องมือธุรกิจจากตลาด รวมถึงการบำรุงดูแลรักษาคลังข้อมูลและการบริหารความเสี่ยง

หนังสือคลังข้อมูลเพื่อการจัดการข้อมูลสมัยใหม่เล่มนี้ได้เรียบเรียงจากเอกสาร ตำรา และการศึกษาค้นคว้าเนื้อหาเพิ่มเติมจากแหล่งข้อมูลต่างๆ รวมถึงการสอดแทรกความรู้จากประสบการณ์จากผู้เขียนเอง เพื่อให้เนื้อหาที่ถ่ายทอดมีความทันสมัยและสอดคล้องกับสถานการณ์โลกปัจจุบัน หนังสือเล่มนี้เหมาะสำหรับอาจารย์ นักวิจัย และนักศึกษาที่เกี่ยวข้องกับสาขาเทคโนโลยีสารสนเทศ วิทยาการคอมพิวเตอร์ วิศวกรรมคอมพิวเตอร์ และคอมพิวเตอร์ธุรกิจ ซึ่งผู้เขียนหวังว่าหนังสือเล่มนี้จะเป็นประโยชน์แก่ผู้ที่สนใจในการบริหารจัดการข้อมูลขนาดใหญ่เพื่อพัฒนาเป็นคลังข้อมูลต่อไป

รองศาสตราจารย์ ดร.รัฐสิทธิ์ สุขะหุต

สารบัญ

คำนิยม	I
คำนำ (พิมพ์ครั้งที่ 1)	II
คำนำ (พิมพ์ครั้งที่ 2)	III
สารบัญ	IV
สารบัญภาพ	XIV
สารบัญตาราง	XVIII
บทที่ 1 แนวคิดของการจัดการข้อมูลสมัยใหม่	1
1.1 วัตถุประสงค์	2
1.2 บทนำ	2
1.3 พัฒนาการข้อมูลในปัจจุบัน	4
1.4 แนวคิดของการจัดการข้อมูลสมัยใหม่	6
1.5 ข้อมูลและข้อมูลสารสนเทศ	9
ประเภทข้อมูล	10
สถาปัตยกรรมข้อมูลสารสนเทศ	11
มุมมองผู้บริหารต่อข้อมูลสารสนเทศ	13
1.6 คลังข้อมูลและข้อมูลสารสนเทศ	14
ฐานข้อมูลปฏิบัติการ	15
การประมวลผลข้อมูลระดับทรานแซคชันแบบออนไลน์	16
ระบบฐานข้อมูลเชิงสัมพันธ์	18
การประมวลผลออนไลน์เชิงวิเคราะห์	20
1.7 การเตรียมความพร้อมสำหรับโครงการคลังข้อมูล	23
แนวคิดการเชื่อมโยงข้อมูลธุรกิจ	23
การวางแผนการดำเนินงาน	24
การสื่อสารกับคนในองค์กร	24
ภาวะข้อมูลท่วมท้น	25
ข้อมูลสกปรก (dirty data)	25

1.8	สรุป	27
1.9	คำถามท้ายบท	28
บทที่ 2	คุณลักษณะสำคัญของคลังข้อมูล	29
2.1	วัตถุประสงค์	30
2.2	บทนำ	30
2.3	คลังข้อมูล (data warehouse)	31
	ประวัติคลังข้อมูล	32
	แนวคิดของการพัฒนาคลังข้อมูล	33
2.4	โครงสร้างพื้นฐานของระบบเครือข่ายและเครื่องแม่ข่าย	34
2.5	คุณสมบัติสำคัญของคลังข้อมูล	35
	Subject-oriented data	35
	Integrated data	36
	Time-variant data	37
	Non-volatile data	37
2.6	สถาปัตยกรรมคลังข้อมูล	38
2.7	ความละเอียดของข้อมูล (data granularity)	39
2.8	ความแตกต่างระหว่างคลังข้อมูลและดาต้ามาร์ท	42
2.9	แนวทางการพัฒนาคลังข้อมูลสำหรับองค์กร	44
	แบบบนลงล่าง Top-down approach	44
	แบบล่างขึ้นบน Bottom-up approach	45
	องค์ประกอบแหล่งข้อมูล	45
	องค์ประกอบพื้นที่พักข้อมูล	47
	องค์ประกอบพื้นที่จัดเก็บข้อมูล	49
2.10	การใช้ประโยชน์จากคลังข้อมูลในเชิงสารสนเทศ	50

สารบัญ

2.11 เมตาดาต้า	52
เมตาดาต้าในระดับปฏิบัติการ	52
เมตาดาต้าสำหรับการสกัดและการแปลงข้อมูล	52
เมตาดาต้าในระดับผู้ใช้งานข้อมูล	52
2.12 สรุป	53
2.13 คำถามท้ายบท	53
บทที่ 3 การวิเคราะห์ความต้องการทางธุรกิจ	55
3.1 วัตถุประสงค์	56
3.2 บทนำ	56
3.3 มุมมองข้อมูลเชิงมิติ	57
ธรรมชาติมิติข้อมูลทางธุรกิจ	58
การวิเคราะห์มิติข้อมูลในเชิงธุรกิจ	61
มิติข้อมูลแบบลำดับขั้น	64
3.4 การกำหนดเป้าหมายของการพัฒนาค้างข้อมูล	65
3.5 ความน่าเชื่อถือของข้อมูล	67
3.6 การจัดเก็บความต้องการของผู้ใช้งาน	68
3.7 การประมวลผลความต้องการเพื่อนำไปสู่การออกแบบคลังข้อมูล	71
3.8 สรุป	73
3.9 คำถามท้ายบท	74
บทที่ 4 แบบจำลองข้อมูลหลายมิติ	75
4.1 วัตถุประสงค์	76
4.2 บทนำ	76
4.3 แบบจำลองข้อมูลเชิงมิติ	76

4.4 การออกแบบแบบจำลองข้อมูลเชิงมิติ	78
การวิเคราะห์แบบจำลองเชิงมิติ	78
การสอบถามข้อมูลเชิงมิติ	80
4.5 โครงสร้างสกีมาของโมเดลเชิงมิติ	82
โครงสร้างแบบดาว	82
โครงสร้างแบบเกล็ดหิมะ	85
โครงสร้างแบบกาแล็กซี	86
4.6 แนวคิดการออกแบบเชิงการวิเคราะห์เชิงมิติ	88
4.7 คุณสมบัติของไดเมนชัน	90
4.8 คุณสมบัติของแฟก	91
ตารางแฟกที่ไม่มีค่าคำนวณ	92
ความละเอียดของข้อมูล	92
การกำหนดคีย์ตัวแทน	93
4.9 มิติเวลา	94
4.10 เมตาดาต้า	97
4.11 มุมมองข้อมูลเชิงสรุป	98
4.12 การออกแบบคลังข้อมูลตามแนวคิดของ Ralph Kimball	98
4.13 ประสิทธิภาพการทำงานของคลังข้อมูล	103
การสร้างดัชนีข้อมูล	104
การแบ่งพาดิชันของฐานข้อมูล	104
การสร้างข้อมูลแบบจัดกลุ่ม	105
การสร้างวิวแบบสรุป	105
4.14 สรุป	106
4.15 คำถามท้ายบท	106

สารบัญ

บทที่ 5	กระบวนการสกัด-แปลงและการโหลดข้อมูล	109
5.1	วัตถุประสงค์	110
5.2	บทนำ	110
5.3	กระบวนการสกัด แปลง และโหลดข้อมูล	111
5.4	ความท้าทายของกระบวนการสกัดและการแปลงข้อมูล	112
5.5	การสกัดข้อมูล	113
	ขั้นตอนการสกัดข้อมูล	113
	ขั้นตอนสำคัญในการสกัดข้อมูล	114
	เทคนิควิธีการการสกัดข้อมูล	115
	ความถี่ในการสกัดข้อมูล	116
5.6	การแปลงข้อมูล	117
	คุณภาพของข้อมูล	117
	ความถูกต้องในค่าต่างๆ ที่เป็นไปได้ของข้อมูล	118
	ความถูกต้องจากความสัมพันธ์ของข้อมูล	120
	การทำความสะอาดข้อมูล	121
	ข้อมูลที่มีความผิดพลาด	123
5.7	การโหลดข้อมูล	124
	การปรับปรุงคลังข้อมูลให้เป็นปัจจุบัน	125
	การโหลดข้อมูลตามช่วงเวลา	125
5.8	วงจรการทำงานของ ETL	127
5.9	สรุป	129
5.10	คำถามท้ายบท	129
 บทที่ 6	 พื้นที่พักข้อมูล	 131
6.1	วัตถุประสงค์	132
6.2	บทนำ	132

6.3	สถาปัตยกรรมของพื้นที่พักข้อมูล	132
6.4	การทำงานของพื้นที่พักข้อมูล	136
6.5	การทำหน้าที่เป็นที่สำรองข้อมูลชั่วคราว	136
6.6	กิจกรรมที่เกิดขึ้นภายในพื้นที่พักข้อมูล	137
	การเชื่อมโยงข้อมูล	138
	การปรับข้อมูล	138
	การลดความขัดแย้งข้อมูล	138
	การตรวจสอบช่วงเวลาของการประมวลผลข้อมูล	139
	การตรวจสอบการเปลี่ยนแปลงข้อมูล	139
	การทำความสะอาดข้อมูล	139
	การประมวลผลข้อมูลสรุปเบื้องต้น	140
	การจัดเก็บข้อมูลที่เก่าในอดีต	140
6.7	สรุป	140
6.8	คำถามท้ายบท	141
บทที่ 7	การประมวลผลออนไลน์เชิงวิเคราะห์	143
7.1	วัตถุประสงค์	144
7.2	บทนำ	144
7.3	แนวคิดการประมวลผลออนไลน์เชิงวิเคราะห์	144
7.4	การประมวลผลโอแลป	146
	การ Roll-up	148
	การ Drill-down	149
	การ Slice and dice	150
	การ Pivot	151
7.5	สรุป	152
7.6	คำถามท้ายบท	153

สารบัญ

บทที่ 8	การออกแบบทางกายภาพ	155
8.1	วัตถุประสงค์	156
8.2	บทนำ	156
8.3	การออกแบบทางกายภาพ	156
	มาตรฐานการพัฒนา	157
	จัดทำแผนโครงการ	157
	การกำหนดพาร์ติชันโครงสร้างข้อมูล	158
	การกำหนดตัวเลือกการจัดกลุ่ม	158
	การจัดทำดัชนี	159
8.4	องค์ประกอบที่ต้องพิจารณาในการออกแบบ	159
	ส่วนประกอบของแบบจำลองทางกายภาพ	161
	การจัดเก็บข้อมูลทางกายภาพ	164
	พื้นที่จัดเก็บโครงสร้างข้อมูล	165
	การจัดทำดัชนีคลังข้อมูล	165
8.5	สรุป	170
8.6	คำถามท้ายบท	170
 บทที่ 9	 ภาษาเอ็มดีเอ็กซ์	 171
9.1	วัตถุประสงค์	172
9.2	บทนำ	172
9.3	ภาษาเอ็มดีเอ็กซ์	172
	ฟังก์ชัน MEMBERS	175
	ฟังก์ชัน CHILDRENS	176
	ฟังก์ชัน DESCENDANTS	177
	ฟังก์ชัน ADDCALCULATEDMEMBERS	177
9.4	การสร้างคำสั่งแบบ Complex query	178

9.5	การกำหนดเงื่อนไขสำหรับการคิวรี	178
	การสร้างค่าคำนวณใหม่	179
	ฟังก์ชัน CURRENTMEMBER และ PARENT	181
	ฟังก์ชัน FIRSTCHILD, LASTCHILD และ CURRENTMEMBER	181
9.6	การท่องตามโครงสร้างแบบลำดับขั้น	182
	ฟังก์ชัน ANCESTER	182
	ฟังก์ชัน EXCEPT	182
	ฟังก์ชัน GENERATE	183
	ฟังก์ชัน LEAD	184
9.7	ฟังก์ชันที่ทำงานกับช่วงเวลา	184
9.8	การกรองข้อมูล (filter) และการเรียงลำดับ (sort)	186
	ฟังก์ชัน FILTER	187
	ฟังก์ชัน ORDER	187
9.9	การวิเคราะห์ผลจากการจัดอันดับสูงสุดและต่ำสุด	188
9.10	ฟังก์ชันสำหรับการคำนวณและเงื่อนไข	189
9.11	สรุป	191
9.12	คำถามท้ายบท	192
บทที่ 10	เหมืองข้อมูล	193
10.1	วัตถุประสงค์	194
10.2	บทนำ	194
10.3	แนวคิดของการทำเหมืองข้อมูล	194
10.4	เหมืองข้อมูลคืออะไร	195
10.5	หลักการทำเหมืองข้อมูล	197
10.6	การเตรียมความพร้อมก่อนการทำเหมืองข้อมูล	198
10.7	การประมวลผลข้อมูลเชิงมิติ OLAP กับเหมืองข้อมูล	201

สารบัญ

10.8	เหมืองข้อมูลและคลังข้อมูล	204
10.9	เทคนิคการทำเหมืองข้อมูล	205
	เทคนิคการจัดกลุ่ม	206
	เทคนิคต้นไม้ตัดสินใจ	209
	การหาเหตุผลจากหน่วยความจำ	213
	การวิเคราะห์การเชื่อมโยง	216
	กฎความสัมพันธ์	217
10.10	ประโยชน์ที่ได้รับจากการทำเหมืองข้อมูล	222
10.11	สรุป	225
10.12	คำถามท้ายบท	225
บทที่ 11	ธุรกิจอัจฉริยะ	227
11.1	วัตถุประสงค์	228
11.2	บทนำ	228
11.3	แนวคิดธุรกิจอัจฉริยะ	228
11.4	องค์ประกอบที่สำคัญของเครื่องมือธุรกิจอัจฉริยะ	229
11.5	การวิเคราะห์ข้อมูลแบบมุมมองหลายมิติ	230
11.6	รายงานในเครื่องมือธุรกิจอัจฉริยะ	233
11.7	สรุป	238
11.8	คำถามท้ายบท	238
บทที่ 12	การบำรุงดูแลรักษาค้างข้อมูล	239
12.1	วัตถุประสงค์	240
12.2	บทนำ	240
12.3	แนวคิดของการบำรุงรักษาระบบคลังข้อมูล	240
	บทบาทและความรับผิดชอบต่อการพัฒนาค้างข้อมูล	241
	การรักษาความปลอดภัยข้อมูล	243

สารบัญ

12.4	ทรัพยากรฮาร์ดแวร์ที่เกี่ยวข้องกับคลังข้อมูล	244
12.5	เทคโนโลยีเวอร์ชวลไลเซชัน	244
12.6	การสำรองและการกู้คืนข้อมูล	247
12.7	การล้างข้อมูล	248
12.8	การบริหารจัดการความเสี่ยง	249
12.9	สรุป	251
12.10	คำถามท้ายบท	251
บรรณานุกรม		253
ดัชนี		254

สารบัญภาพ

รูปที่ 1.1	พัฒนาการของเทคโนโลยีเว็บ	3
รูปที่ 1.2	ตัวอย่างเอเจนท์ที่ใช้ในแต่ละอุตสาหกรรม	3
รูปที่ 1.3	เครือข่ายการสื่อสารที่เชื่อมหากันบนโลกออนไลน์	4
รูปที่ 1.4	สถาปัตยกรรมข้อมูลสารสนเทศ	12
รูปที่ 1.5	โครงสร้างระบบ SAP อีอาร์พี	18
รูปที่ 1.6	โครงสร้างฐานข้อมูลร้านเช่าวิดีโอ ในระดับ conceptual design	20
รูปที่ 2.1	ความแตกต่างระหว่างฐานข้อมูลปฏิบัติการคลังข้อมูล	33
รูปที่ 2.2	การรวบรวมข้อมูลจากแหล่งข้อมูลต่างๆ เป็นคลังข้อมูล	36
รูปที่ 2.3	สถาปัตยกรรมโดยรวมของระบบคลังข้อมูล	38
รูปที่ 2.4	ระดับความละเอียดข้อมูลที่จัดเก็บภายในคลังข้อมูล และทิศทางการไหลของข้อมูล	40
รูปที่ 2.5	โครงสร้างสถาปัตยกรรมของมีติเอเตอร์	48
รูปที่ 2.6	การไหลของข้อมูลสู่คลังข้อมูลตามช่วงเวลา	50
รูปที่ 3.1	แสดงภาพโครงสร้างองค์กรในฝ่ายการตลาด โดยแบ่งออกเป็นลำดับขั้น	59
รูปที่ 3.2	ไฮเปอร์คิวบ์แสดงมิติของข้อมูลทางธุรกิจ	62
รูปที่ 3.3	มิติของข้อมูลในแต่ละประเภทของธุรกิจ	63
รูปที่ 3.4	ตัวอย่าง Information Packages สำหรับข้อมูลการขาย	72
รูปที่ 4.1	ตัวอย่างแบบจำลองข้อมูลเชิงมิติ (dimension model)	77
รูปที่ 4.2	แสดงขั้นตอนจากความต้องการสู่การออกแบบโมเดล ข้อมูลเชิงมิติ	78
รูปที่ 4.3	แบบจำลองข้อมูลสำหรับ Automaker sales	79
รูปที่ 4.4	โมเดลข้อมูลที่มีโครงสร้างแบบดาว (star schema)	83
รูปที่ 4.5	โมเดลข้อมูลที่มีโครงสร้างแบบเกล็ดหิมะ (snowflake schema)	85
รูปที่ 4.6	โมเดลข้อมูลที่มีโครงสร้างแบบกาแล็กซี (galaxy schema)	87

สารบัญภาพ

รูปที่ 4.7	การเชื่อมโยงข้อมูลระหว่างแฟก Enrollment และ Time table ในโครงสร้างกาแล็กซี	87
รูปที่ 4.8	โครงสร้างข้อมูลเชิงมิติสำหรับห้างสรรพสินค้า	89
รูปที่ 4.9	ตัวอย่างตารางแฟกที่ไม่มีค่าคำนวณ	92
รูปที่ 4.10	การสร้างคีย์ตัวแทน CustomerKey สำหรับข้อมูลลูกค้า	94
รูปที่ 4.11	ตัวอย่างตารางไทม์เมนชันสำหรับมิติเวลา (time dimension)	95
รูปที่ 4.12	แสดงความสัมพันธ์ระหว่างตารางไทม์เมนชันและตารางแฟก	100
รูปที่ 5.1	การออกแบบขั้นตอนการทำ ETL	111
รูปที่ 5.2	ความสัมพันธ์ข้อมูลสายการบินที่สัมพันธ์กับเวลา	116
รูปที่ 5.3	การออกแบบขั้นตอนการทำ ETL	125
รูปที่ 5.4	แสดงวงจรการทำงานของ ETL	128
รูปที่ 6.1	สถาปัตยกรรมของพื้นที่พักข้อมูล	133
รูปที่ 6.2	การเตรียมการก่อนการนำเข้าข้อมูล	137
รูปที่ 7.1	คิวบ์แสดงมุมมองการวิเคราะห์ข้อมูลแบบมิติไทม์เมนชัน	145
รูปที่ 7.2	คิวบ์แสดงมุมมองการวิเคราะห์ข้อมูลแบบมิติไทม์เมนชัน	146
รูปที่ 7.3	การสร้างมุมมองข้อมูลในลักษณะของการ Roll-up	148
รูปที่ 7.4	การสร้างมุมมองข้อมูลในลักษณะของการ Drill-down	149
รูปที่ 7.5	การสร้างมุมมองข้อมูลในลักษณะของการ Slice	150
รูปที่ 7.6	การสร้างมุมมองข้อมูลในลักษณะของการ Dice	151
รูปที่ 7.7	การสร้างมุมมองข้อมูลในลักษณะของการ Pivot	152
รูปที่ 8.1	ขั้นตอนการออกแบบทางกายภาพ	157
รูปที่ 8.2	กระบวนการแปลงแบบจำลองทางตรรกะเป็นแบบจำลองทางกายภาพ	161
รูปที่ 8.3	องค์ประกอบทางกายภาพของคลังข้อมูล	161
รูปที่ 8.4	ตัวอย่างนิยามโครงสร้างใน SQL	162
รูปที่ 8.5	แบบจำลองทางตรรกะและแบบจำลองทางกายภาพ	163

สารบัญภาพ

รูปที่ 8.6	โครงสร้างข้อมูลในคลังข้อมูล	165
รูปที่ 8.7	ตัวอย่างดัชนี B-Tree	166
รูปที่ 8.8	ตัวอย่างดัชนีบีตแมป	168
รูปที่ 8.9	การสืบค้นข้อมูลจากดัชนีบีตแมป	168
รูปที่ 9.1	ไดเมนชันโมเดลสำหรับ Sales	174
รูปที่ 10.1	สนับสนุนการตัดสินใจดำเนินการทำเหมืองข้อมูล	196
รูปที่ 10.2	การค้นหาข้อมูลที่มีคุณค่า	198
รูปที่ 10.3	ขั้นตอนการค้นหาความรู้	199
รูปที่ 10.4	กระบวนการค้นหาความรู้	201
รูปที่ 10.5	ความแตกต่างระหว่าง OLAP และ Data mining	202
รูปที่ 10.6	การทำเหมืองข้อมูลในสภาพแวดล้อมของคลังข้อมูล	205
รูปที่ 10.7	การกระจายตัวของข้อมูลลูกค้าตามอายุและรายได้เฉลี่ยของลูกค้าทั้งหมด 100 คน	208
รูปที่ 10.8	ผลลัพธ์จากการแบ่งกลุ่มลูกค้าทั้งหมด 100 คนออกเป็น 3 กลุ่ม โดยพิจารณาจากข้อมูลอายุและรายได้เฉลี่ย	208
รูปที่ 10.9	ต้นไม้การตัดสินใจสำหรับผู้ซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊ก	209
รูปที่ 10.10	ต้นไม้การตัดสินใจสำหรับการออกไปเล่นกีฬากลางแจ้งโดยพิจารณาจากสภาพอากาศ	213
รูปที่ 10.11	การกระจายของข้อมูลดอกไอริส 3 ชนิด จำนวน 15 ดอก	214
รูปที่ 10.12	การกระจายของข้อมูลดอกไอริส 3 ชนิด จำนวน 15 ดอก และข้อมูลดอกไอริสที่ต้องการทำนายชนิดพร้อมเพื่อนบ้านที่ใกล้ที่สุด 3 ตัว	216
รูปที่ 11.1	โครงสร้างตารางข้อมูลแบบ OLTP	231
รูปที่ 11.2	แบบจำลองโครงสร้างข้อมูลเชิงมิติสำหรับข้อมูลการขายสินค้า	232
รูปที่ 11.3	ตัวอย่างเครื่องมือ BI ที่แสดงผลรายงานแบบตาราง	234
รูปที่ 11.4	ตัวอย่างเครื่องมือ BI ที่แสดงผลรายงานแบบแผนภูมิ	234

สารบัญภาพ

รูปที่ 11.5	ตัวอย่างเครื่องมือ BI ที่แสดงผลรายงานแบบตัวชี้วัด	235
รูปที่ 11.6	การแสดงผล Dashboard ภาพรวมสารสนเทศ	236
รูปที่ 11.7	ตัวอย่างรายงานของ Prelytis ที่สนับสนุนการทำงานแบบ Multi-platform	237
รูปที่ 12.1	กิจกรรมที่เกี่ยวข้องกับการดูแลรักษาคลังข้อมูล	243
รูปที่ 12.2	สถาปัตยกรรมของเครื่องแม่ข่ายเสมือนที่ใช้อุปกรณ์การจัดเก็บร่วมกัน	245
รูปที่ 12.3	การสำรองข้อมูลแบบ Incremental backup	248

สารบัญตาราง

ตารางที่ 1.1	ความแตกต่างระหว่างข้อมูลปฏิบัติและคลังข้อมูล	22
ตารางที่ 9.1	รายละเอียดของตารางไดเมนชันสำหรับคิวบ์ Sales	173
ตารางที่ 9.2	รายละเอียดของตารางแฟก Sales	174
ตารางที่ 10.1	OLAP จะใช้ในการวิเคราะห์ข้อมูลที่ผ่านมา และเหมืองข้อมูลสำหรับการทำนายอนาคต	203
ตารางที่ 10.2	ความแตกต่างของ OLAP และ Data mining	203
ตารางที่ 10.3	การประยุกต์ใช้เหมืองข้อมูลเพื่อการวิเคราะห์ ในแต่ละลักษณะงาน	206
ตารางที่ 10.4	ข้อมูลสภาพอากาศ	211
ตารางที่ 10.5	แสดงระยะห่างระหว่างข้อมูลดอกไอริส	215
ตารางที่ 10.6	แสดงสินค้าที่ลูกค้าจำนวน 10 คน ซื้อจากร้านค้าแห่งหนึ่ง	218
ตารางที่ 10.7	แสดงเซตรายการที่ประกอบด้วยแอทริบิวต์ 2 แอทริบิวต์และค่าสนับสนุนของแต่ละเซตรายการ	219
ตารางที่ 10.8	กฎความสัมพันธ์ที่สร้างจากเซตรายการที่ประกอบด้วย 2 แอทริบิวต์และค่าความเชื่อมั่น	222
ตารางที่ 10.9	แนวทางการประยุกต์เทคนิคเหมืองข้อมูล	223

ปัจจุบันในยุคสมัยของโลกสังคมดิจิทัลและสื่อออนไลน์ที่มีการพึ่งพาเทคโนโลยีสารสนเทศมากยิ่งขึ้น วิทยาการข้อมูลได้เข้ามามีบทบาทสำคัญ องค์กรต่างๆ ทั้งหน่วยงานในภาครัฐและเอกชนที่มีการจัดเก็บข้อมูลในรูปของสื่ออิเล็กทรอนิกส์ต่างให้ความสำคัญต่อการจัดเก็บข้อมูลเพื่อนำข้อมูลมาใช้ในการประกอบการวางแผน วิเคราะห์ และเป็นข้อมูลสนับสนุนการตัดสินใจของผู้บริหาร แนวคิดของการจัดการข้อมูลสมัยใหม่เป็นศาสตร์ที่ว่าด้วยการบริหารจัดการข้อมูลอย่างเป็นระบบ เป็นการผสมผสานระหว่างวิทยาการคอมพิวเตอร์ หลักสถิติ คณิตศาสตร์ และซอฟต์แวร์ และต้องอาศัยความรู้ทางด้านการจัดการแบบบูรณาการ ซึ่งจำเป็นต้องให้ความสำคัญต่อการออกแบบโครงสร้างและสถาปัตยกรรมข้อมูลที่ดี เพื่อให้การใช้ข้อมูลสารสนเทศเกิดประโยชน์สูงสุดต่อองค์กร

DATA WAREHOUSING
FOR MODERN DATA
MANAGEMENT



CHIANG MAI
UNIVERSITY PRESS

ISBN 978-616-398-310-7



9 786163 983107

ราคา 450 บาท

01

บทที่ 1

แนวคิดของการจัดการข้อมูล
สมัยใหม่

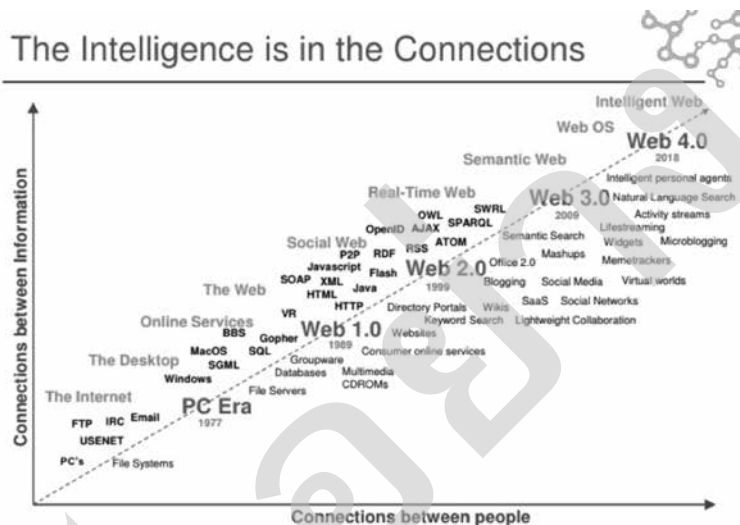
1.1 วัตถุประสงค์

- เพื่ออธิบายถึงแนวคิดของการบริหารจัดการข้อมูลสมัยใหม่
- เพื่ออธิบายโครงสร้างของข้อมูลเพื่อนำไปสู่ข้อมูลสารสนเทศ
- เพื่อให้เห็นความแตกต่างระหว่างฐานข้อมูลปฏิบัติการและคลังข้อมูล
- เพื่ออธิบายถึงแนวทางการใช้ประโยชน์ข้อมูลในเชิงการวิเคราะห์และการตัดสินใจสำหรับผู้บริหาร

1.2 บทนำ

ปัจจุบันเทคโนโลยีสารสนเทศมีบทบาทสำคัญเป็นอย่างมากในองค์กรธุรกิจ ทั้งภาครัฐและเอกชน เทคโนโลยีถูกนำมาใช้เป็นเครื่องมือในการเพิ่มประสิทธิภาพการทำงาน การบริหาร เพื่อสนับสนุนให้องค์กรบรรลุเป้าหมายและพันธกิจขององค์กร อีกทั้งยังเป็นการเพิ่มขีดความสามารถในการพัฒนาและเป็นการสร้างความได้เปรียบเหนือคู่แข่ง จากการเติบโตของเทคโนโลยีสารสนเทศและการสื่อสารที่มีพัฒนาการเพิ่มขึ้นอย่างรวดเร็ว ทำให้เกิดการสื่อสารแลกเปลี่ยนข้อมูล โดยเฉพาะเทคโนโลยีสมาร์ทโฟน (smart phone) เกิดเป็นแอปพลิเคชันบนระบบเครือข่ายอย่างมากมาย การเกิดชุมชนแบบออนไลน์และเครือข่ายสังคมออนไลน์ (social media) ทำให้มีปริมาณข้อมูลที่เพิ่มมากขึ้น ประกอบกับความหลากหลายของข้อมูลที่ต้องดำเนินการมีปริมาณที่เพิ่มขึ้นอย่างรวดเร็ว ก่อให้เกิดข้อมูลขนาดใหญ่ที่อยู่บนโลกออนไลน์ที่ไร้ข้อจำกัดในเรื่องของเวลาและสถานที่ ระบบการสืบค้นข้อมูลออนไลน์หรือเสิร์ชเอนจิน (search engine) เป็นเครื่องมือที่ได้รับความนิยม ทำให้การเข้าถึงข้อมูลที่อยู่บนโลกอินเทอร์เน็ตทำได้โดยง่าย โดยจากรูปที่ 1.1 แสดงถึงพัฒนาการเริ่มต้นจากการจัดเก็บและการเผยแพร่ข้อมูลที่มีการจัดเก็บข้อมูลในรูปแบบของไฟล์หรือฐานข้อมูลอยู่ในคอมพิวเตอร์แบบตั้งโต๊ะ (PC era) จนถึงการเผยแพร่ข้อมูลข่าวสารผ่านทางเว็บไซต์ (world wide web - web 1.0) การสื่อสารในโลกสังคมออนไลน์ (the social web - web 2.0) การค้นคืนข้อมูลเชิงความหมาย (the semantic web - web 3.0) จนถึงปัจจุบันอยู่ในยุคสมัยของเว็บชาญฉลาด (the intelligence web - web 4.0) ซึ่งมีทิศทางการพัฒนาเว็บในยุคถัดไปที่ข้อมูลบนโลกออนไลน์มีความเชื่อมโยงกันมากขึ้นในลักษณะของเครือข่ายเชิงความหมาย นำไปสู่การพัฒนาซอฟต์แวร์แอปพลิเคชันที่มีความชาญฉลาด โดยอาศัยเทคโนโลยีการประมวลผลทางภาษาธรรมชาติ (Natural Language Processing: NLP) ที่คอมพิวเตอร์จะมีความสามารถในการเข้าใจภาษามนุษย์ในเชิงความหมายมากขึ้น การแบ่งปันข้อมูลข่าวสารหรือโซเชี่ยลบุ๊กมาร์ก (human social tagging – “folksonomies”) เพื่อการจัดเก็บ แบ่งหมวดหมู่ และการค้นหาข้อมูลบนโลกออนไลน์มีความสะดวกมากยิ่งขึ้น และรวมถึงแนวคิดการจัดเก็บข้อมูลโดยอาศัยแนวคิดออนโทโลยี (ontology) เพื่ออธิบายความสัมพันธ์โครงสร้างความรู้แบบลำดับชั้นเชิงวัตถุ (object hierarchical data structure) และการนำเสนอความรู้ (knowledge representation)

language) นอกจากนี้ในยุคของ web 4.0 แอปพลิเคชันที่มีความฉลาดบนโลกออนไลน์ประเภทเอเจนต์ที่มีความฉลาด (Intelligent Agent: IA) จะทำให้เกิดการสื่อสารระหว่าง application-to-application ในลักษณะของการสื่อสารระหว่างเอเจนต์ด้วยตัวเอง เพื่อแลกเปลี่ยนข้อมูลข่าวสารโดยไม่มีมนุษย์มาเกี่ยวข้อง ตัวอย่างเอเจนต์ในปัจจุบันได้แก่ การแลกเปลี่ยนข้อมูลบนเครือข่ายเว็บ การตรวจสอบราคาหุ้น การพยากรณ์สภาพอากาศ การเปรียบเทียบราคาโรงแรมที่พักในประเทศต่างๆ ดังแสดงในรูปที่ 1.2



รูปที่ 1.1 พัฒนาการของเทคโนโลยีเว็บ

ที่มา: <http://www.slideshare.net/novaspivack/web-evolution-nova-spivack-twine> (2009)

Agent	Percepts	Actions	Goals	Environment
ALVINN	Images, Position	Steering, Control speed	Drive from A to B	Roads, Vehicles, Hazards
Google engine	Web pages, Search queries	Navigate web, Gather info.	Find best search matches	Internet
Deep Blue	Current board state	Choose next move	Win the game!	Opponent, Game board
Financial Forecaster	Stock market data	Pick stocks to buy & sell	Maximize investments	Stock market, company reports
Medical Diagnostic	Symptoms, Test results	Tests, Treatments	Healthy patient	Patient, Hospital

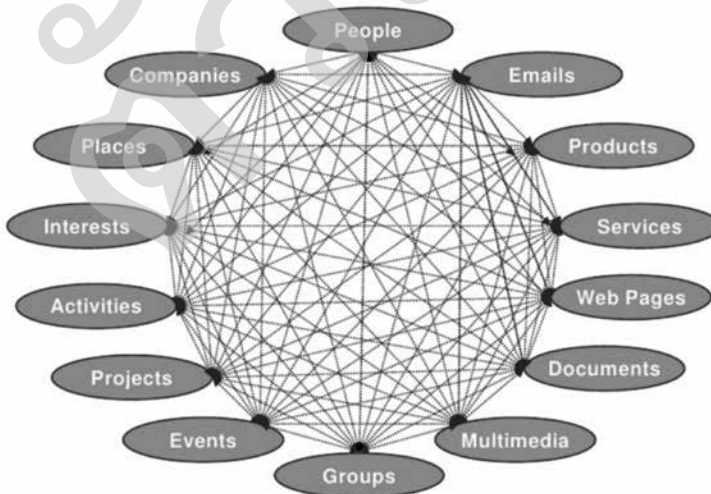
รูปที่ 1.2 ตัวอย่างเอเจนต์ที่ใช้ในแต่ละอุตสาหกรรม

ที่มา: http://pages.cs.wisc.edu/~bsettles/cs540/lectures/02_agents.pdf (2009)

การบริหารจัดการข้อมูลแบบบูรณาการภายใต้สภาพแวดล้อมและการเปลี่ยนแปลงในยุคโลกาภิวัตน์ ที่เทคโนโลยีอินเทอร์เน็ตมีบทบาทสำคัญในการสนับสนุนการทำงาน การทำพาณิชย์อิเล็กทรอนิกส์ การทำธุรกรรมอิเล็กทรอนิกส์ผ่านทางระบบเครือข่ายอินเทอร์เน็ต อาทิ การซื้อขายสินค้าแบบออนไลน์ ทำให้เกิดการสื่อสารแบบไร้พรมแดน ปัจจุบันจะพบว่าข้อมูลที่จัดเก็บภายในองค์กรมีปริมาณเพิ่มมากขึ้น รูปแบบการใช้ประโยชน์จากข้อมูลที่จัดเก็บไม่เพียงแต่มีการพัฒนาเป็นฐานข้อมูลเพื่อรองรับการทำงาน แต่จะต้องมองถึงการใช้ประโยชน์เชิงสารสนเทศ การวางแผนและประกอบ การตัดสินใจ ที่ต้องออกแบบโครงสร้างของข้อมูลที่จัดเก็บ เรียกว่าเป็นสถาปัตยกรรมโครงสร้างข้อมูลที่ต้องมีการวางแผน ซึ่งการวางแผนเป็นภาพรวมของระบบการบริหารจัดการข้อมูลตั้งแต่เริ่มต้น

1.3 พัฒนาการข้อมูลในปัจจุบัน

ปัจจุบันพัฒนาการเทคโนโลยีฮาร์ดแวร์มีขนาดที่เล็กลงเป็นอย่างมากเมื่อเทียบกับเทคโนโลยีในช่วง 10-20 ปีที่ผ่านมา แต่กลับมีประสิทธิภาพเพิ่มมากขึ้นแบบก้าวกระโดด อุปกรณ์ฮาร์ดแวร์มีความสามารถในการประมวลผล สามารถรองรับการทำงานแบบหลายๆ งานพร้อมกัน (multi-tasking) ระบบปฏิบัติการและซอฟต์แวร์ใหม่ๆ และพื้นที่จัดเก็บที่รองรับข้อมูลขนาดใหญ่ นอกจากนี้ความก้าวหน้าของเทคโนโลยีเครือข่ายและการสื่อสารที่ทำให้ข้อมูลเกิดการเชื่อมหากันแบบออนไลน์ ดังแสดงในรูปที่ 1.3 ปรากฏจากข้อจำกัดในเรื่องของสถานที่และเวลา ข้อมูลมีการสร้างขึ้นมาใหม่อยู่ตลอดเวลาทั้งในโลกธุรกิจและในโลกของสังคมออนไลน์ การเปลี่ยนแปลงสำคัญคือ ผู้ใช้งานทุกคนกลายมาเป็นผู้สร้างข้อมูล (content creator) ซึ่งแต่เดิมจะถูกจำกัดอยู่แค่ผู้ดูแลระบบข้อมูล หรือผู้ดูแลเว็บไซต์ เท่านั้น



รูปที่ 1.3 เครือข่ายการสื่อสารที่เชื่อมหากันบนโลกออนไลน์

ที่มา: <http://www.slideshare.net/novaspivack/web-evolution-nova-spivack-twine> (2009)

เทคโนโลยีอุปกรณ์ชาญฉลาด (smart device) ได้เพิ่มความสะดวกในการเข้าถึงระบบเครือข่ายและสารสนเทศ อุปกรณ์อิเล็กทรอนิกส์และอุปกรณ์สมองกลแบบฝังตัว สามารถเป็นตัวกำเนิดข้อมูลได้อย่างอัตโนมัติ ทำให้ปริมาณข้อมูลเพิ่มมากขึ้นอย่างรวดเร็ว ตัวอย่างเช่น อุปกรณ์เซ็นเซอร์ที่ใช้สำหรับตรวจวัดค่าต่างๆ เครื่องจับเวลา เครื่องวัดอุณหภูมิ เครื่องบันทึกเวลา กล้องวงจรปิด และนาฬิกาอัจฉริยะ (smart watch) ฯลฯ ซึ่งอุปกรณ์เหล่านี้ทำหน้าที่เสมือนเป็นเซ็นเซอร์ที่ตรวจจับและบันทึกผล โดยจะทำการสร้างข้อมูลขึ้นมาใหม่ตลอดเวลา เกิดเป็นแนวคิดใหม่เรียกว่า “อินเทอร์เน็ตของสรรพสิ่ง” (Internet of Things: IOT) ปัจจุบันอุปกรณ์เหล่านี้วันจะมีปริมาณที่เพิ่มมากขึ้นเรื่อยๆ และมีบทบาทสำคัญใกล้ตัวเราเข้ามาเรื่อยๆ ดังตัวอย่างเช่น อุปกรณ์อิเล็กทรอนิกส์แบบพกพาที่สามารถตรวจจับการเต้นของหัวใจ ระยะทางการก้าวเดินหรือวิ่ง หรือแม้แต่รูปแบบของการนอน และมีการส่งข้อมูลเหล่านี้ผ่านทางระบบเครือข่ายเพื่อนำข้อมูลมาประมวลผล แสดงให้เห็นถึงพฤติกรรมการกิน การนอน หรือการออกกำลังกายของคนๆ นั้น

ปัจจุบันองค์กรเริ่มมีการจัดเก็บข้อมูลสำคัญในรูปของเอกสารอิเล็กทรอนิกส์ ซึ่งแนวโน้มของข้อมูลขนาดใหญ่ในวันจะมีปริมาณที่เพิ่มมากขึ้น สาเหตุหนึ่งเกิดจากความสามารถของเทคโนโลยีที่มีความทันสมัยในการสร้างข้อมูลเดิมที่เริ่มต้นจากระดับองค์กรหรือหน่วยงาน แต่ปัจจุบันความสามารถในการสร้างข้อมูลในระดับบุคคลมีทั้งจากอุปกรณ์โทรศัพท์มือถือ กล้องถ่ายรูป กล้องวิดีโอ และข้อมูลที่มีการแบ่งปันบนโลกออนไลน์ที่อยู่ในรูปของเครือข่ายสังคมออนไลน์ วิดีโอ และบล็อกส่วนตัว เป็นต้น ซึ่งขนาดของข้อมูลที่อยู่บนโลกออนไลน์นั้นนับตั้งแต่ปี ค.ศ. 1980 จนถึงปี ค.ศ. 2012 มีข้อมูลที่ถูกรวบรวมประมาณ 2.5 เอ็กซาไบต์ (2.5×10^{18}) และในปี ค.ศ. 2014 จะมีข้อมูลมากถึง 2.3 เซตทราไบต์ (2.3×10^{21})¹ ซึ่งถือได้ว่าเป็นข้อมูลที่มากมายมหาศาลที่อยู่บนระบบเครือข่าย และหากถามหาความเป็นเจ้าของข้อมูลอาจจะไม่สามารถระบุชี้ชัดได้อย่างชัดเจน

นอกจากนี้มาตรฐานการเชื่อมต่อระบบเครือข่ายอินเทอร์เน็ตในปัจจุบันที่ใช้อินเทอร์เน็ตโพรโทคอล หรือไอพี (Internet Protocol : IP)² ซึ่งได้ปรับจากมาตรฐาน IPv4 หรือไอพีในเวอร์ชัน 4 ซึ่งมีข้อจำกัดในเรื่องของการจัดสรรคลาสของอินเทอร์เน็ตทำให้หมายเลขไอพีที่เป็นสาธารณะ (public IP address) ที่มีขนาดเพียง 32 บิต ไม่เพียงพอต่อการใช้งาน และจากจำนวนอุปกรณ์อิเล็กทรอนิกส์ที่ต้องการใช้หมายเลขไอพี โดยได้มีการพัฒนาไอพีเป็นเวอร์ชัน 6 หรือ IPv6 ทำให้สามารถมีจำนวนของหมายเลขไอพีที่เพิ่มขึ้นถึง 128 บิต และทำให้หมายเลขไอพีที่จะเกิดขึ้นในอนาคตนั้นมีจำนวนเท่ากับ 3.4×10^{38} หรือมากกว่าหมายเลขที่อยู่ปัจจุบันถึง 7.9×10^{28} เท่า หรือสามารถให้หมายเลขไอพีได้ถึง 4.3 ล้านล้านหมายเลข และทำให้มีหมายเลขไอพีที่มากพอที่จะกำหนดให้แก่อุปกรณ์อิเล็กทรอนิกส์อื่นๆ ที่นอกเหนือจากเครื่องคอมพิวเตอร์ของเราเอง ลองจินตนาการว่าต่อจากนี้ไปอุปกรณ์อิเล็กทรอนิกส์

¹“Big Data at the Speed of Business - What is big data?”. www.ibm.com. Retrieved 2014-11-23.

²เมื่อวันที่ 6 มิถุนายน ค.ศ. 2016 ได้กำหนดให้เป็นวันเริ่มต้นการใช้งาน IPv6 อย่างเป็นทางการ เรียกว่า World IPv6 Day โดย IPv6 ได้ผ่านการทดสอบการใช้งานมาจนถึงในระดับที่สามารถนำใช้กับอุปกรณ์สินค้าและบริการต่างๆ ได้

และเครื่องใช้ไฟฟ้าภายในบ้าน เช่น โทรศัพท์ ตู้เย็น หรือไมโครเวฟ สามารถมีหมายเลขไอพีกำกับในแต่ละอุปกรณ์ และสามารถเชื่อมต่ออุปกรณ์เหล่านั้นผ่านทางระบบเครือข่ายเน็ตเวิร์ก ซึ่งจะทำให้สามารถสื่อสารหรือสั่งการให้อุปกรณ์เหล่านั้นทำงานผ่านทางระบบเครือข่าย เช่นทำการเปิดเครื่องไฟฟ้า หรือเพิ่ม/ลดอุณหภูมิของแอร์ภายในบ้านจากที่ทำงาน หรือผ่านทางโทรศัพท์มือถือได้ ซึ่งหากอุปกรณ์เหล่านี้สามารถรับ-ส่งข้อมูลผ่านทางเครือข่ายได้ จะก่อให้เกิดเป็นข้อมูลมหาศาลบนระบบอินเทอร์เน็ต

1.4 แนวคิดของการจัดการข้อมูลสมัยใหม่

แนวคิดของการจัดการข้อมูลสมัยใหม่เป็นแนวคิดการใช้เทคโนโลยีสารสนเทศและข้อมูลข่าวสาร เพื่อเชื่อมโยงกับพันธกิจและเป้าหมายในเชิงยุทธศาสตร์ขององค์กร โดยมีกระบวนการจัดการเทคโนโลยีสารสนเทศ รวมถึงการจัดการข้อมูลที่ทันสมัยที่สนับสนุนการดำเนินงานเชิงยุทธศาสตร์ที่สอดคล้องกับการเปลี่ยนแปลงของระบบเศรษฐกิจและสภาวะแวดล้อมของการแข่งขันในปัจจุบัน เทคโนโลยีสารสนเทศถูกนำมาใช้เพื่อเป็นเครื่องมือในการเพิ่มประสิทธิภาพการทำงานในระดับปฏิบัติการ และในขณะเดียวกันมุ่งผลสัมฤทธิ์ที่เป็นเป้าหมายเชิงกลยุทธ์ขององค์กร การจัดการข้อมูลเป็นพื้นฐานสำคัญของการจัดการข้อมูลสมัยใหม่ ระบบข้อมูลที่ดีย่อมต้องรองรับการปฏิบัติงานประจำ สามารถเชื่อมโยงการทำงานทั้งระดับการปฏิบัติการและระดับบริหาร มีกลไกการจัดเก็บรวบรวมข้อมูล รวมถึงการตรวจสอบความถูกต้องของข้อมูลที่เชื่อมโยงกระบวนการทำงานต่างๆ อย่างสอดคล้องกัน

วิทยาการข้อมูล (data science) เป็นวิทยาการสำหรับการจัดการความรู้สมัยใหม่ที่ผสมผสานศาสตร์จากหลายสาขาทั้งด้านวิทยาศาสตร์ คณิตศาสตร์ประยุกต์ สถิติ การจัดการความรู้ วิทยาการคอมพิวเตอร์ ความรู้ในด้านการบริหารและการจัดการ และความรู้เฉพาะสาขา ซึ่งนักวิทยาการข้อมูลจะกลายมาเป็นบุคคลที่มีบทบาทสำคัญในการใช้ความรู้ทั้งในด้านการบริหารและวิชาการโดยใช้ข้อมูลที่มีอยู่เพื่อสร้างเป็นองค์ความรู้ใหม่ สิ่งที่เป็นความท้าทายคือ การประยุกต์ทฤษฎีทางด้านวิทยาการข้อมูล เพื่อตอบโจทย์และปัญหาทางธุรกิจ องค์กรสมัยใหม่จะมีการเชื่อมโยงกระบวนการทำงานภายในองค์กรในแต่ละแผนก/ฝ่ายกับการจัดการข้อมูล ผู้เชี่ยวชาญด้านการจัดการข้อมูลจะต้องทำงานร่วมกับฝ่ายต่างๆ ควบคู่ไปกับผู้บริหาร และระดับปฏิบัติการ เพื่อที่จะแก้ปัญหากระบวนการทำงานภายในผสมผสานกับวิทยาการข้อมูลในการจัดการความรู้สมัยใหม่

การจัดการกับข้อมูลขนาดใหญ่เป็นเรื่องที่มีความท้าทายสำหรับโลกปัจจุบัน เนื่องจากเทคโนโลยีสารสนเทศด้านการสื่อสารมีพัฒนาการและการเปลี่ยนแปลงที่รวดเร็วแบบก้าวกระโดด ข้อมูลใหม่เกิดขึ้นอยู่ตลอดเวลา ข้อมูลในปัจจุบันมีความหลากหลายของรูปแบบ (heterogeneous data) มากยิ่งขึ้น ทำให้มีความเป็นพลวัตของข้อมูลสูงเกิดเป็นข้อมูลขนาดใหญ่ (big data) ที่มีความซับซ้อนทำให้เกิดความท้าทายในการบริหารจัดการ รวมถึงการจัดเก็บและการนำเอาข้อมูลมาใช้ประโยชน์ใน

ลักษณะของการควิรีเพื่อการสืบค้นข้อมูล การนำข้อมูลมาใช้ในการวิเคราะห์ ซึ่งจะเห็นได้ว่าการจัดการข้อมูลสมัยใหม่นั้น มุ่งเน้นทางด้านประสิทธิภาพของการวิเคราะห์และการประมวลผลข้อมูลที่มีปริมาณมหาศาล เป็นประเด็นที่มีความท้าทาย ที่ดึงดูดความสนใจจากสถาบันการศึกษา อุตสาหกรรม ทั้งในภาครัฐบาลและเอกชน โดยมีการนำเสนอแนวคิดในการจัดการข้อมูล และรวมถึงเทคนิคต่างๆ เพื่อใช้สำหรับการประมวลผลข้อมูลจากระบบและการประยุกต์ใช้ข้อมูลในด้านต่างๆ สำหรับมุมมองของการจัดการข้อมูลขนาดใหญ่จะอาศัยกลไกการจัดเก็บข้อมูล (cloud storage) และการประมวลผลแบบก้อนเมฆ (cloud computing)

การประมวลผลแบบก้อนเมฆ เป็นโมเดลการจัดเก็บข้อมูลที่มีอยู่มหาศาล โดยข้อมูลที่ถูกจัดเก็บในรูปของสื่ออิเล็กทรอนิกส์จะมองเสมือนเป็นข้อมูลที่จัดรวมอยู่ในที่เดียวกัน แต่ในทางกายภาพข้อมูลอาจถูกจัดเก็บแบบกระจายได้หลายเครื่องแม่ข่าย และเครื่องแม่ข่ายอาจจะอยู่ในหลายศูนย์ข้อมูล ภายในสถานที่เดียวกัน แม้แต่ปัจจุบันศูนย์ข้อมูลที่ให้บริการพื้นที่แบบคลาวด์ตั้งอยู่ในพื้นที่ต่างประเทศ ทำให้การเข้าถึงเพื่อใช้งานปราศจากข้อจำกัดในทางกายภาพ การเข้าถึงเพื่อใช้งาน และการบริหารจัดการ ผู้ให้บริการพื้นที่จัดเก็บข้อมูลบนคลาวด์มีหน้าที่รับผิดชอบด้านความเสถียรภาพของข้อมูล อำนาจความสะดวกในการเชื่อมโยง กำหนดสิทธิ์และจำกัดสิทธิ์การใช้งานและการเข้าถึงแก่ผู้ใช้งานโดยผ่านช่องทางกรให้บริการเว็บ (web application programming interface) การใช้บริการคลาวด์ยังเป็นการลดความเสี่ยงด้านสถาปัตยกรรมเครือข่ายที่อาจจะเกิดขึ้นกับองค์กร เนื่องจากองค์กรสามารถตรวจสอบระดับการให้บริการ (Service Level Agreement: SLA) จากผู้ให้บริการคลาวด์ ซึ่งหากเปรียบเทียบข้อดีและข้อเสียของการใช้บริการคลาวด์ และการดูแลรักษาระบบต่างๆ เหล่านี้ภายในองค์กรเอง จะพบว่าในองค์กรอาจพบความเสี่ยงมากกว่าหากดูแลรักษาอุปกรณ์เหล่านี้เอง นอกจากนี้เทคโนโลยีแบบคลาวด์ยังช่วยลดข้อจำกัดของความรู้ทางด้านเทคนิคของพนักงานในองค์กร ทำให้องค์กรสามารถขยายขีดความสามารถในการแข่งขันด้านเทคโนโลยีสารสนเทศโดยไม่มีข้อจำกัดแนวคิดการบริการแบบคลาวด์ยังได้ครอบคลุมไปถึงสถาปัตยกรรมแบบก้อนเมฆ แพลตฟอร์มของระบบคอมพิวเตอร์ และรวมถึงรูปแบบการจัดเก็บข้อมูลเพื่อให้เกิดการทำงานแบบผสมผสานระหว่างเทคโนโลยีสารสนเทศ อุปกรณ์สำหรับการเชื่อมต่อเครือข่ายสังคม ข้อมูลมัลติมีเดีย อุปกรณ์สมองกลแบบฝังตัว เซ็นเซอร์ RFID และกลไกการประมวลผลข้อมูล โดยการให้บริการแบบคลาวด์ (cloud services) ถูกแบ่งออกเป็น 5 ลักษณะ ดังต่อไปนี้

- Infrastructure-as-a-Service (IaaS) – เป็นการให้บริการที่เป็นเวอร์ชวลแมชีนที่สามารถเข้าถึงได้ผ่านเครือข่าย รองรับความต้องการใช้งานในการประมวลผลหรือพื้นที่จัดเก็บ
- Platform-as-a-Service (PaaS) – บริการด้านแพลตฟอร์มสำหรับซอฟต์แวร์ (เช่น เว็บแอปพลิเคชัน ดาต้าเบสเซิร์ฟเวอร์ ระบบประมวลผลกลางสำหรับองค์กรขนาดใหญ่ และมิดเดิลแวร์อื่นๆ โดยทำงานภายใต้การควบคุมด้านความปลอดภัยสูง) ที่เรียกใช้งานได้ผ่านเว็บแอปพลิเคชัน

- Software-as-a-Service (SaaS) – เป็นบริการด้านแอปพลิเคชันโดยคิดค่าบริการเป็นไลเซนส์ของผู้ใช้ หรือตามปริมาณการใช้งาน
- Data-as-a-Service (DaaS) – ให้บริการข้อมูลและข้อมูลสารสนเทศจากการเชื่อมโยงพื้นที่สำหรับจัดเก็บข้อมูลแบบคลาวด์ที่เป็นแหล่งเก็บข้อมูลดิบหรือข้อมูลเพื่อถูกจัดเตรียมไว้สำหรับการวิเคราะห์
- Business Process-as-a-Service (BPaaS) – เป็นคลาวด์สำหรับบริการด้านธุรกิจที่ต้องการปรับปรุงกระบวนการทางธุรกิจและผลลัพธ์ทางธุรกิจได้

ยกตัวอย่างกรณีศึกษาต่อไปนี้เป็นสำหรับปัจจุบันองค์กรต่างๆ ได้มีการพัฒนาศูนย์ข้อมูลหรือดาต้าเซ็นเตอร์ (data center) ของตนเองขึ้นมา เพื่อใช้สำหรับจัดเก็บข้อมูลและการพัฒนาเป็นซอฟต์แวร์ระบบเพื่อใช้ภายในองค์กร โดยมีการจัดซื้อจัดหาอุปกรณ์เครื่องแม่ข่าย สตอเรจสำหรับจัดเก็บข้อมูล อุปกรณ์เชื่อมต่อกับระบบเครือข่าย รวมถึงระบบการรักษาความปลอดภัย ระบบป้องกันการบุกรุก (intrusion prevention) ซึ่งภายในห้องดาต้าเซ็นเตอร์ยังต้องมีระบบไฟฟ้า เครื่องสำรองไฟ (UPS) หรือระบบไฟฟ้าสำรองเครื่องกำเนิดไฟฟ้า ระบบการรักษาความปลอดภัย เช่น อุปกรณ์สแกนลายนิ้วมือ และกล้องวงจรปิด ซึ่งอุปกรณ์เหล่านี้จำเป็นต้องมีเจ้าหน้าที่ที่มีความรู้ความสามารถในแต่ละด้าน ทั้งด้านเทคนิคเครื่องแม่ข่าย เน็ตเวิร์กและนักพัฒนาเทคโนโลยีสารสนเทศ โดยพบว่าเมื่อเวลาผ่านไปอุปกรณ์ต่างๆ เหล่านี้จะต้องมีการบำรุงดูแลรักษา ซึ่งต้องใช้เงินงบประมาณจำนวนมาก อีกทั้งการบริหารจัดการบุคคลที่จะดูแลระบบเหล่านี้ให้มีความเสถียรภาพและประสิทธิภาพ การใช้บริการเครื่องแม่ข่ายที่อยู่บนคลาวด์ ทำให้องค์กรสามารถมีระบบที่ให้บริการ ไม่ว่าจะเป็นระบบปฏิบัติการเครื่องแม่ข่าย และระบบการป้องกันรักษาความปลอดภัยที่จัดหาโดยผู้ให้บริการแบบคลาวด์ โดยที่มีเงื่อนไขของการรับบริการ ทำให้ลดปัญหาในเรื่องของการบริหารจัดการระบบต่างๆ ดังกล่าวลงไปได้

1.5 ข้อมูลและข้อมูลสารสนเทศ

ข้อมูล (data) หมายถึงข้อมูลที่จัดเก็บในรูปแบบข้อมูลอิเล็กทรอนิกส์โดยมีวิธีการบันทึกผ่านซอฟต์แวร์หรือแอปพลิเคชันสำเร็จรูป ข้อมูลเกิดจากกระบวนการทางธุรกรรม (transaction) จากการปฏิบัติงานประจำที่มีการบันทึกตามกระบวนการทางธุรกิจ ตัวอย่างเช่น การสั่งซื้อสินค้าของลูกค้าและข้อมูลการขายสินค้า ธุรกรรมทำให้เกิดการเปลี่ยนแปลงข้อมูล เช่น การเพิ่ม ลบ และการแก้ไข การดำเนินงานทางธุรกิจมักจะเกี่ยวข้องกับเจ้าหน้าที่ที่ปฏิบัติงานในแต่ละแผนก แต่ละฝ่าย ที่เป็นงานประจำ เจ้าหน้าที่บัญชีทำการบันทึกข้อมูลบัญชี เจ้าหน้าที่การเงินจะจัดทำรายงานงบประมาณ และเจ้าหน้าที่งานพัสดุก็ทำหน้าที่ในการจัดซื้อ จัดจ้าง การสั่งซื้อสินค้าและข้อมูลการตรวจรับสินค้า รวมถึงการจัดทะเบียนวัสดุ-ครุภัณฑ์ที่ได้มีการจัดซื้อ ในมุมมองของสารสนเทศ ข้อมูลดังกล่าวอาจเรียกได้ว่าเป็นข้อมูลดิบ (raw data) เนื่องจากเป็นข้อมูลที่ยังไม่ผ่านการตรวจสอบ การกลั่นกรอง หรือการประมวลผล ยังไม่สามารถนำมาใช้สำหรับประกอบการตัดสินใจโดยที่ยังไม่มีความหมายใดๆ ในเชิงการบริหาร

ข้อมูลสารสนเทศ (information) หมายถึงข้อมูลที่ผ่านการวิเคราะห์ สังเคราะห์ การประมวลผล เพื่อให้เป็นข้อมูลที่มีความถูกต้อง มีความน่าเชื่อถือ เป็นข้อมูลที่สามารถนำไปใช้ประกอบการวิเคราะห์ ประกอบการวางแผนและการตัดสินใจ ข้อมูลสารสนเทศเป็นข้อมูลที่ได้จากการจัดเก็บรวบรวมมาจากข้อมูลปฏิบัติการ และเป็นข้อมูลที่ถูกนำมาประมวลผลให้อยู่ในรูปแบบที่เหมาะสมต่อการนำไปใช้ประโยชน์หรือการตัดสินใจของผู้บริหาร เช่น การนำข้อมูลในระดับเรคคอร์ดมาจัดทำให้อยู่ในรูปแบบข้อมูลเชิงสรุป เพื่อนำมาแสดงผลในรูปแบบของตารางหรือแผนภูมิแบบต่างๆ ข้อมูลสารสนเทศที่ดีควรมีคุณสมบัติที่สำคัญทั้ง 5 ด้าน ดังต่อไปนี้

1. ข้อมูลมีความถูกต้อง (data correction) ข้อมูลจะต้องมีความน่าเชื่อถือ ขั้นตอนการตรวจสอบและการยืนยันความถูกต้อง สามารถอ้างอิงถึงแหล่งที่มาของข้อมูลได้
2. ความซ้ำซ้อนของข้อมูลที่จัดเก็บ (data redundancy) ข้อมูลที่จัดเก็บจะต้องไม่มีความซ้ำซ้อนของข้อมูล หมายถึงข้อมูลในแต่ละเรื่องจะต้องมีการเก็บเพียงชุดเดียว หากเป็นระบบฐานข้อมูลก็ควรจะต้องมีเพียงฐานเดียว
3. การใช้ประโยชน์ข้อมูลร่วมกัน (data sharing) ข้อมูลที่จัดเก็บในองค์กรจะต้องมีความสามารถในการใช้ประโยชน์ข้อมูลร่วมกัน การยอมให้มีการเข้าถึงข้อมูลโดยบุคคลหรือฐานข้อมูลที่มีการเชื่อมโยงกัน
4. ความปลอดภัยของข้อมูล (data security) การกำหนดระดับขั้นของความปลอดภัย รวมถึงการกำหนดสิทธิ์การเข้าถึงของผู้ใช้งาน หมายถึงยอมให้เข้าถึงข้อมูลได้โดยผู้ที่รับผิดชอบหรือผู้ที่เกี่ยวข้องเท่านั้น
5. ความสอดคล้องต่อการนำไปใช้ประโยชน์ (data usability) การจัดเก็บข้อมูลควรเป็นข้อมูลที่จำเป็น และมีความเหมาะสมต่อลักษณะการนำไปใช้ประโยชน์

กรณีศึกษาตัวอย่าง ประโยชน์ของการใช้ข้อมูลสารสนเทศเพื่อการตัดสินใจของบริษัทผลิตชิ้นส่วนอะไหล่ยนต์

บริษัทผลิตชิ้นส่วนอะไหล่ยนต์เพื่อส่งขายทั้งในและต่างประเทศ ที่ประกอบกิจการมานานกว่า 30 ปี ซึ่งในระยะ 4-5 ปีหลัง ผู้บริหารได้เริ่มให้ความสำคัญกับข้อมูลสารสนเทศ สาเหตุมาจากผลประกอบการที่ลดลง ปัญหาการรั่วไหลที่ไม่สามารถระบุได้ว่าเกิดจากสาเหตุใด การตัดสินใจไม่อยู่บนพื้นฐานของข้อมูลที่เป็นจริงและเป็นปัจจุบัน การแก้ไขปัญหาทางบริษัทจึงได้มีการรวบรวมข้อมูลต่างๆ ของบริษัทโดยพัฒนาเป็นฐานข้อมูลของบริษัท และพัฒนาเป็นรายงานข้อมูลสารสนเทศในภาพรวม เพื่อให้ผู้บริหารสามารถมองเห็นภาพรวมของการดำเนินงานของกิจการ สามารถรู้ถึงสถานะปัจจุบันและความเคลื่อนไหว เช่น สถานะทางการเงิน และผลประกอบการของบริษัท อีกทั้งยังสามารถมองภาพย่อยลงไปในระดับของโครงการ ทั้งกำไร-ขาดทุนในแต่ละโครงการ ต้นทุน สถานะการดำเนินงานในแต่ละขั้นตอน รวมถึงสามารถตัดสินใจและสั่งการบนพื้นฐานของข้อมูล ตัวอย่างเช่น การสั่งซื้อวัตถุดิบตามปริมาณความต้องการของลูกค้า และการวางแผนการผลิตตามกำลังการผลิตของเครื่องจักร และพนักงานที่มีอยู่จากการสั่งซื้อของลูกค้า เป็นต้น

ประเภทข้อมูล

การจัดเก็บข้อมูลอิเล็กทรอนิกส์ในแต่ละองค์กรมักจะพบว่าข้อมูลที่เก็บอยู่ในปัจจุบันนั้นมีความหลากหลาย (heterogeneous data) หมายถึงข้อมูลจัดเก็บมีรูปแบบที่แตกต่างกัน องค์กรหนึ่งๆ อาจมีข้อมูลได้ในหลายรูปแบบ ข้อมูลอาจอยู่ในรูปของไฟล์ข้อความ (text file) หรือไบนารี (binary) ข้อมูลแบบกึ่งโครงสร้าง ข้อมูลมีมิติเดียว เช่น ภาพ เสียง และภาพเคลื่อนไหว ข้อมูลบางประเภทมีการบันทึกโดยซอฟต์แวร์ที่พัฒนาขึ้นมา ซึ่งจะมีรูปแบบข้อมูลที่จัดเก็บโดยเฉพาะ แต่สำหรับในบางองค์กรมีการวางแผนการจัดเก็บข้อมูลอย่างเป็นระบบ เช่น การจัดเก็บข้อมูลในรูปของฐานข้อมูลเชิงสัมพันธ์ ทำให้ข้อมูลที่จัดเก็บเป็นไปตามมาตรฐานและมีความสัมพันธ์ที่เชื่อมโยงกัน ซึ่งความหลากหลายของข้อมูลเหล่านี้เป็นประเด็นท้าทายสำหรับนักออกแบบข้อมูล

การจำแนกประเภทของข้อมูลสามารถแบ่งออกได้ตามโครงสร้างของข้อมูลที่จัดเก็บ โดยสามารถแบ่งออกได้เป็น 3 ประเภทหลักๆ ได้แก่

- 1) **ข้อมูลที่มีโครงสร้าง (Structured data)** ได้แก่ข้อมูลที่มีการวิเคราะห์ ออกแบบอย่างเป็นระบบ การจัดเก็บข้อมูลมีโครงสร้างข้อมูลที่มีรูปแบบที่เป็นมาตรฐานเดียวกัน เช่น ข้อมูลที่อยู่ในรูปของฐานข้อมูลเชิงสัมพันธ์ที่มีการจัดเก็บข้อมูลในรูปของตารางประกอบด้วยแถว (row) และสดมภ์ (column)
- 2) **ข้อมูลแบบกึ่งโครงสร้าง (Semi-structured data)** เป็นข้อมูลที่มีการกำหนดโครงสร้างข้อมูลแบบหยาบๆ เช่นการกำหนดรูปแบบข้อมูล (formatting) แต่ไม่ได้กำหนดในระดับของเนื้อหา ตัวอย่างของข้อมูลประเภทนี้ได้แก่ข้อมูล SGML³, HTML และ XML