

เอกสารประกอบการสอน

วิชาการออกแบบคลังข้อมูล

DATA WAREHOUSE DESIGN



อาจารย์โกเมศ อัมพวัน

คณะวิทยาการสารสนเทศ

มหาวิทยาลัยบูรพา

นิยามและส่วนประกอบของคลังข้อมูล



- 2.1 แผนการสอนประจำบท
- 2.2 บทนำ
- 2.3 คุณลักษณะของข้อมูลในคลังข้อมูล
- 2.4 ส่วนประกอบของคลังข้อมูล
- 2.5 คลังข้อมูลและดาต้ามาร์ท
- 2.6 วิธีการสร้างคลังข้อมูล
- 2.7 แนวปฏิบัติสำหรับการสร้างคลังข้อมูล
- 2.8 คำถามท้ายบท



แผนการสอนประจำบท

วัตถุประสงค์ของบทเรียน

- ทบทวนคำจำกัดความของคลังข้อมูล
- อธิบายเกี่ยวกับคุณลักษณะต่างๆ ของคลังข้อมูล
- การแยกความแตกต่างระหว่างคลังข้อมูลและดาต้ามาร์ท (Data marts)
- ศึกษาเกี่ยวกับส่วนประกอบของคลังข้อมูล
- ศึกษาเกี่ยวกับวิธีการสร้างคลังข้อมูล

เนื้อหาของบทเรียน

เนื้อหาในบทนี้จะประกอบด้วยคุณลักษณะของคลังข้อมูล ส่วนประกอบของคลังข้อมูล นิยามของดาต้ามาร์ท ความแตกต่างของคลังข้อมูลและดาต้ามาร์ท วิธีการสร้างคลังข้อมูล และแนวปฏิบัติในการสร้างคลังข้อมูล

อุปกรณ์ที่ใช้ในการเรียน-การสอน

- เอกสารประกอบการสอน
- เครื่องคอมพิวเตอร์
- เครื่องฉายภาพสไลด์

กิจกรรมการเรียน-การสอน

- อธิบายพร้อมยกตัวอย่างประกอบ
- ศึกษาจากเอกสารประกอบการสอน
- ฝึกปฏิบัติการตามที่มอบหมาย
- ทำแบบฝึกหัดท้ายบท

การวัดและประเมินผล

- การตอบคำถามระหว่างการเรียน-การสอน
- การทำแบบทดสอบย่อยท้ายบท
- การตรวจงานตามที่มอบหมาย

บทนำ





การทำธุรกิจอย่างชาญฉลาด (**Business intelligence, BI**) จะประกอบไปด้วยการทำงานหลัก 2 ขั้นตอน คือ การเปลี่ยนข้อมูลให้เป็นข้อมูลสารสนเทศ และการเปลี่ยนข้อมูลสารสนเทศไปเป็นองค์ความรู้ โดยการสร้างคลังข้อมูลนั้นจะเป็นส่วนหนึ่งของ BI ที่จะทำให้การเปลี่ยนข้อมูลดิบให้เป็นข้อมูลสารสนเทศที่มีประโยชน์ อาทิ การสร้างข้อมูลที่เป็นผลสรุป การสร้างข้อมูลที่ถูกอธิบายหรือแสดงในเชิงเปรียบเทียบ และอื่นๆ โดยข้อมูลสารสนเทศที่ได้จะสามารถนำไปเป็นเครื่องมือในการประกอบการตัดสินใจในการดำเนินธุรกิจต่างๆ ได้



การทำงานของคลังข้อมูลจะประกอบไปด้วย 4 ฟังก์ชันการทำงานหลักด้วยกัน คือ (1) การเลือกข้อมูลดิบมาบางส่วน โดยทำการเลือกเฉพาะข้อมูลที่สำคัญหรือเป็นข้อมูล que ผู้ใช้สนใจ (2) การประมวลผลกับข้อมูลดิบเหล่านั้น อาทิ การทำให้ข้อมูลต่างๆ เป็นมาตรฐาน การทำให้ข้อมูลมีความถูกต้อง สมบูรณ์ ครบถ้วน เป็นต้น (3) การจัดเก็บข้อมูลที่สำคัญเหล่านั้นไว้ในคลังข้อมูล และ (4) การเรียกใช้งานข้อมูลที่สำคัญที่ถูกจัดเก็บอยู่ในคลังข้อมูล

คุณลักษณะของข้อมูลในคลังข้อมูล

ก่อนที่จะทำการสร้างคลังข้อมูล เราควรจะต้องทราบถึงคุณลักษณะของคลังข้อมูลว่ามีลักษณะเด่น หรือลักษณะพิเศษของคลังข้อมูล เป็นอย่างไร มีแหล่งข้อมูลเป็นอย่างไรบ้าง และปัจจัยอื่นๆ อีกมากมาย นอกจากนั้นเราจำเป็นที่จะต้องพิจารณาถึงส่วนประกอบต่างๆ ที่แตกต่างกันที่จะทำให้การทำงานสามารถตอบสนองความต้องการของผู้ใช้ได้ดีที่สุด รวมถึงการศึกษาเกี่ยวกับคุณลักษณะพื้นฐานของคลังข้อมูลด้วย ซึ่งจากบทที่ผ่านมาจะทำให้เราทราบถึงความต้องการของการสร้างคลังข้อมูล แต่เรายังไม่ทราบว่าคลังข้อมูลคืออะไร? มีลักษณะเป็นอย่างไร? เพื่อให้เข้าใจนิยามพื้นฐานของคลังข้อมูล ลองพิจารณานิยามที่ได้จาก **“Bill Inmon”** ที่ซึ่งเป็นผู้ที่ได้รับการขนานนามว่าเป็น **“บิดาของคลังข้อมูล”** ได้ทำกรนิยามคลังข้อมูลไว้ว่า



อ้างอิงภาพ <http://formacioncontinua.medellin.upb.edu.co/2013/BI/Bill-Inmon.php>

นิยามคลังข้อมูล

“Data warehouse is a collection of data in support of management’s decision that have 4 characteristics : (1) subject-oriented (2) integrated (3) nonvolatile and (4) time variant”

ในส่วนของ “Sean Kelly” ซึ่งเป็น “ผู้ทรงอิทธิพลต่อการสร้างคลังข้อมูล” ก็ได้ให้คำจำกัดความเกี่ยวกับคลังข้อมูลไว้เช่นกัน ดังนี้

“Data in data warehouse is (1) separate (2) available (3) integrated (4) time stamped (5) subject oriented (6) nonvolatile and (7) accessible”

จากนิยามที่ใกล้เคียงกันของบุคคลที่มีชื่อเสียงเกี่ยวกับการสร้างคลังข้อมูลทั้งสอง เราจะสามารถสรุปและทำการอธิบายถึงนิยามของคลังข้อมูลได้ แต่ก่อนที่จะทำการศึกษาถึงนิยามของคลังข้อมูล ลองพิจารณาคำถามเบื้องต้นต่อไปนี้เพื่อที่จะทราบถึงคุณลักษณะหลักของคลังข้อมูล ก่อนที่จะทำความเข้าใจหรือทำการสร้างคลังข้อมูลคุณเกิดข้อสงสัยเหล่านี้หรือไม่

1) ข้อมูลในคลังข้อมูลมีลักษณะเป็นอย่างไร?

2) ข้อมูลในคลังข้อมูลแตกต่างจากข้อมูลในระบบการดำเนินงานอย่างไร?

3) ทำไมข้อมูลจากทั้งสองระบบต้องแตกต่างกัน?

4) เราจะสามารถใช้ข้อมูลในคลังข้อมูลได้อย่างไร?

จากคำถามเหล่านี้ เราลองพิจารณาถึงคุณลักษณะของข้อมูลที่สำคัญในคลังข้อมูลดังนี้

1

ข้อมูลที่ถูกจัดเก็บตามหัวข้อที่สนใจ
(Subject-oriented data)

ก่อนที่จะพิจารณาข้อมูลในคลังข้อมูล ลองพิจารณาข้อมูลในระบบการดำเนินงานซึ่งเป็นข้อมูลที่เราคุ้นเคยเป็นอย่างดี โดยข้อมูลที่ถูกจัดเก็บอยู่ในระบบการดำเนินงานจะถูกจัดเก็บโดยแยกตามแอปพลิเคชันหรือฟังก์ชันการใช้งานต่างๆ อาทิ

- ข้อมูลการสั่งซื้อสินค้าจากลูกค้าที่ประกอบไปด้วยข้อมูลลูกค้า ข้อมูลเกี่ยวกับสินค้า และจำนวนที่ต้องการสั่งซื้อสินค้า
- ข้อมูลคลังสินค้าจะประกอบไปด้วย รหัสสินค้า ชื่อสินค้า หมวดหมู่สินค้า และจำนวนสินค้าคงเหลือในคลังสินค้า
- การตรวจสอบเครดิตของลูกค้าจะเก็บข้อมูล ลูกค้า และธนาคาร เป็นต้น

จากตัวอย่างข้างต้น จะทำให้เราได้เห็นภาพกว้างๆ ของระบบการดำเนินงานของธุรกิจต่างๆ ที่มีการจัดเก็บข้อมูลที่สอดคล้องกับการทำธุรกรรมหนึ่งๆ กับธุรกิจนั้นๆ และยังสอดคล้องกับฟังก์ชันการทำงานต่างๆ แต่ในส่วน of คลังข้อมูลนั้นจะมีการเก็บข้อมูลที่แตกต่างจากระบบดำเนินงาน ข้อมูลในคลังข้อมูลจะถูกเก็บและเชื่อมโยงด้วย “หัวข้อทางธุรกิจ (*Business subject*)” ที่เกี่ยวเนื่องและมีความสำคัญต่อการดำเนินธุรกิจ เช่น “บริษัทผู้ผลิตสินค้า” จะมีความเกี่ยวเนื่องกับข้อมูลการผลิตสินค้า การขายสินค้า การส่งสินค้า การจัดเก็บสินค้าเข้าสู่คลังสินค้า และอื่นๆ ข้อมูลเหล่านี้จะเป็นข้อมูลที่เป็นหัวข้อทางธุรกิจที่สำคัญของบริษัทที่จะส่งผลต่อผลกำไร-ขาดทุนของธุรกิจนั้นๆ ลองพิจารณาอีกตัวอย่างหนึ่ง คือ “บริษัทค้าปลีก” จะมีข้อมูลการขาย ณ จุดขายสินค้าเป็นหัวข้อทางธุรกิจที่สำคัญ เป็นต้น จากตัวอย่างธุรกิจทั้งสอง เราจะสามารถสรุปเกี่ยวกับ “หัวข้อทางธุรกิจ” ได้ว่าเป็น “ฟังก์ชัน กิจกรรม หรือการดำเนินธุรกิจที่สำคัญและส่งผลถึงผลกำไรหรือขาดทุนของบริษัทได้ รวมถึงเป็น ข้อมูลที่สามารถวัดหรือประเมินผลสัมฤทธิ์ได้”

In the data warehouse, data is not stored by operational applications, but by business subjects.

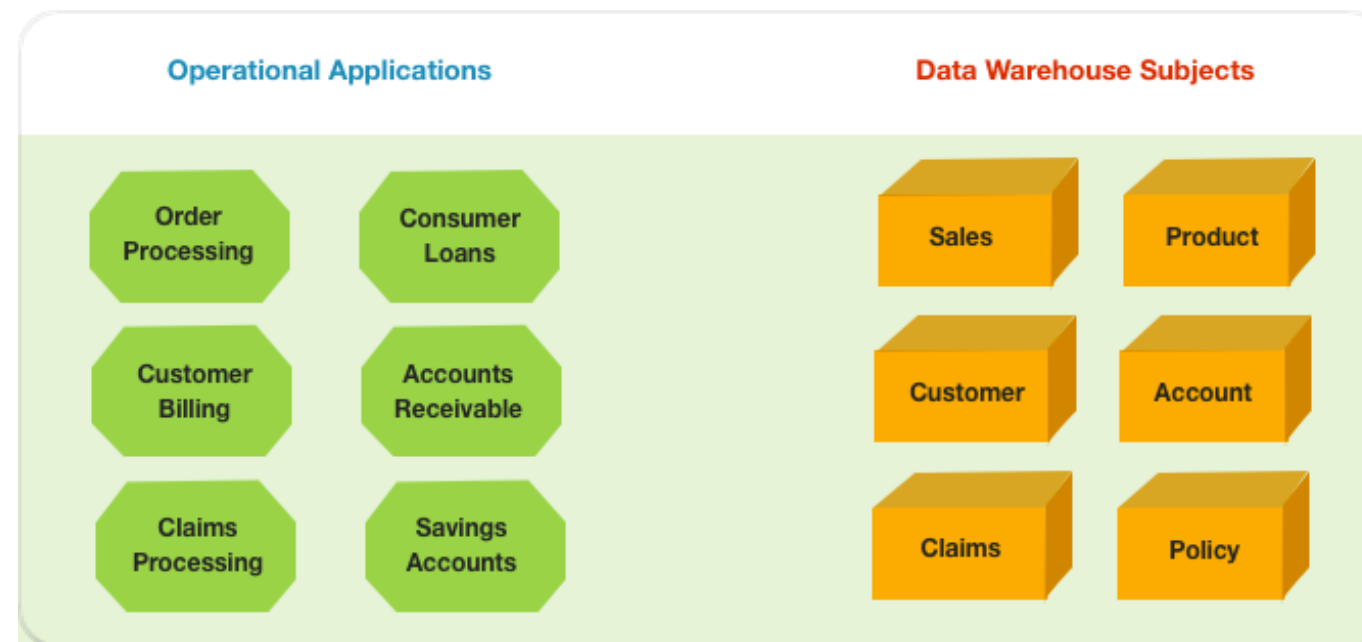
Operational Applications



Data Warehouse Subjects



รูปที่ 2-1 การเปรียบเทียบการจัดเก็บข้อมูลระหว่างระบบการดำเนินงานและคลังข้อมูล



รูปที่ 2-1 การเปรียบเทียบการจัดเก็บข้อมูลระหว่างระบบการดำเนินงานและคลังข้อมูล

ในการที่จะเข้าใจถึงความแตกต่างระหว่างข้อมูลที่ถูกจัดเก็บไว้ในระบบการดำเนินงานและคลังข้อมูล ลองพิจารณารูปที่ 2-1 ที่แสดงการเปรียบเทียบการจัดเก็บข้อมูลจากทั้งสองระบบ โดยจากรูปเราจะสังเกตได้ว่าข้อมูลที่ถูกเก็บไว้ในระบบการดำเนินงานจะถูกเก็บแยกตามแต่ละแอปพลิเคชัน อาทิ ข้อมูลการสั่งซื้อสินค้า ข้อมูลการยืมเงินของลูกค้า ข้อมูลการออกใบเสร็จให้กับลูกค้า ข้อมูลการเรียกเคลมประกัน เป็นต้น ซึ่งจากแอปพลิเคชันต่างๆ เราจะเห็นว่าได้ว่าระบบทั้งสองมีวัตถุประสงค์ต่างกัน โดยระบบการดำเนินงานจะเก็บข้อมูลเพื่อเป็นการลงบันทึกการทำธุรกรรมทางธุรกิจเสียเป็นส่วนใหญ่ แต่การจัดเก็บข้อมูลในคลังข้อมูลจะเป็นการเก็บข้อมูลเพื่อสร้างเป็นข้อมูลเชิงกลยุทธ์สำหรับการประกอบการตัดสินใจในการดำเนินการต่างๆ

เพื่อให้เห็นความแตกต่าง ลองพิจารณาข้อมูลการเรียกเคลมประกันซึ่งจะเป็นข้อมูลที่ถูกจัดเก็บในบริษัทที่ทำธุรกิจประกัน โดยระบบการดำเนินงานจะทำการจัดเก็บข้อมูลการเคลมประกันของลูกค้าแต่ละราย เช่น ณ วันหนึ่งๆ มีลูกค้าที่ซื้อกรมธรรม์หมายเลขใดได้ทำการติดต่อมาเพื่อขอรับการเคลมประกันบ้าง ซึ่งวัตถุประสงค์หลักจะเป็นการเก็บข้อมูลเพื่อเป็นหลักฐานการทำธุรกิจกันระหว่างลูกค้าและบริษัท แต่สำหรับคลังข้อมูลจะเป็นการเก็บข้อมูลการเคลมประกันของลูกค้า โดยจะมุ่งเน้นที่ข้อมูลที่เป็นข้อเท็จจริงและตัวชี้วัดที่เกี่ยวข้องกับหัวข้อนั้นๆ เช่น สาเหตุของการเคลมประกัน จำนวนที่ต้องจ่ายค่าชดเชย และอื่นๆ ข้อมูลเหล่านี้จะเป็นข้อมูลที่เกี่ยวข้องกับการดำเนินธุรกิจโดยตรง เมื่อผู้บริหารทราบถึงค่าชดเชยที่ต้องจ่ายในแต่ละเดือน จะทำให้ทราบถึงผลกำไรและจะสามารถตัดสินใจที่จะคิดหรือดำเนินกลยุทธ์ต่างๆ ที่เพิ่มขึ้นจากเดิมได้

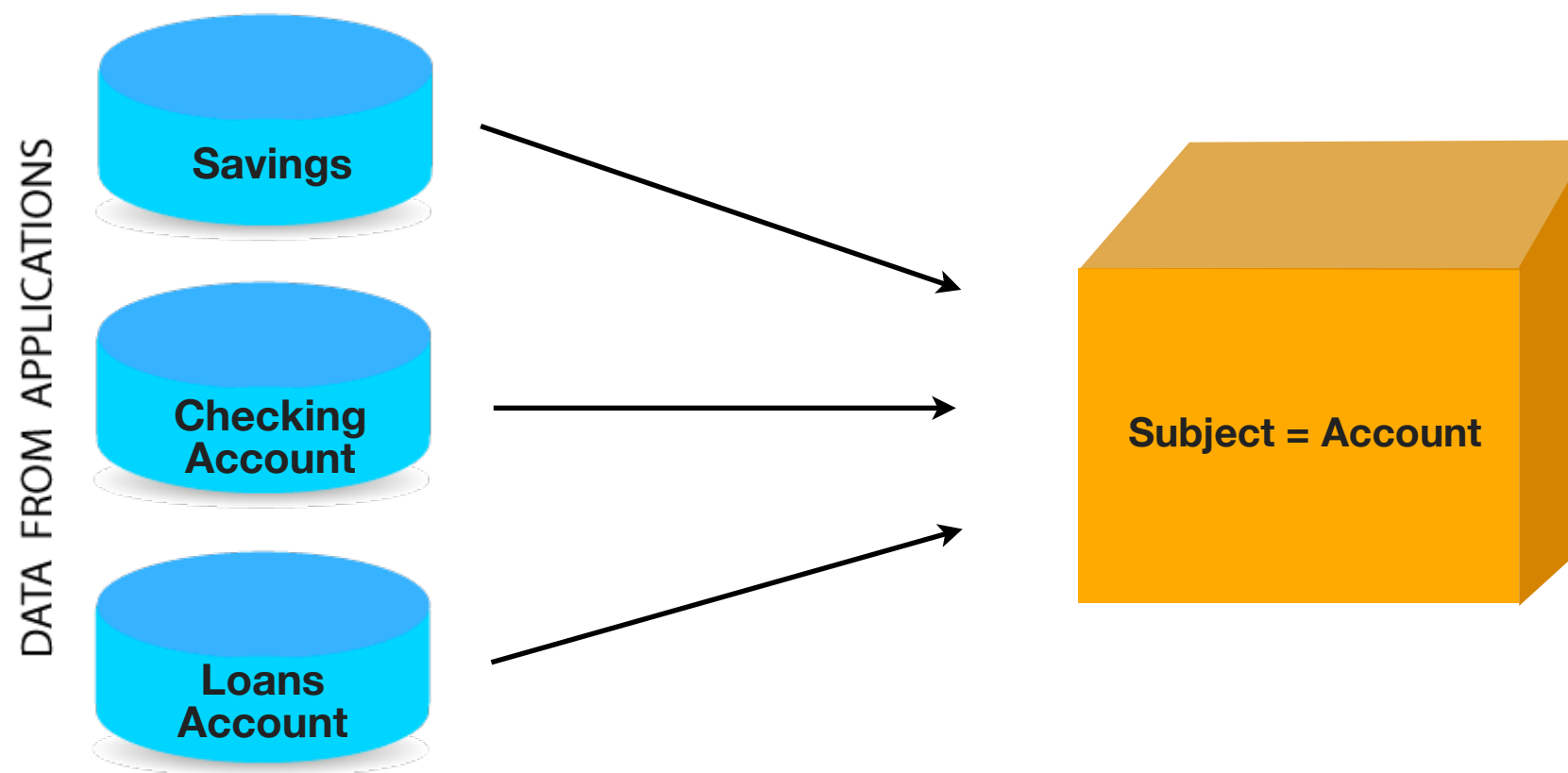


2

ข้อมูลที่ถูกรวมมาจากหลายแหล่งข้อมูล
(Integrated data)

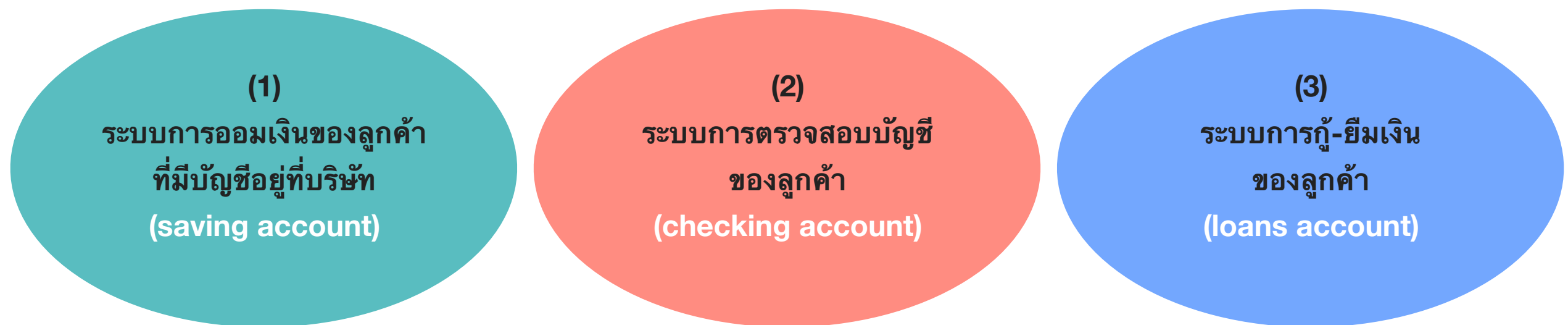
ในการสร้างข้อมูลเชิงกลยุทธ์จากคลังข้อมูลให้มีความครบถ้วนสมบูรณ์ เราอาจจำเป็นต้องเรียกใช้ข้อมูลจากหลายๆ แอปพลิเคชัน (หลายระบบการดำเนินงานหรือหลายแหล่งข้อมูล) โดยข้อมูลที่มาจากหลายระบบอาจมีความแตกต่างกันในเรื่องของระบบการจัดการฐานข้อมูลที่ใช้ (Database management system, DBMS) รูปแบบของแฟ้มข้อมูล (File format) หรือการจัดเก็บข้อมูลส่วนย่อยๆ (Data segment) และแอปพลิเคชันต่างๆ ที่มีความแตกต่างในเรื่องของเค้าโครงของแฟ้มข้อมูล (File layout) การแทนข้อมูลที่เป็นอักขระ (Character code representation) และการตั้งชื่อฟิลด์ต่างๆ ที่สื่อถึงข้อมูลเดียวกัน (Field naming convention) ตัวอย่างเช่น ในการสร้างคลังข้อมูลหนึ่งๆอาจมีการเรียกใช้ข้อมูลจากบริษัท Metro Mail, A.C. Nielsen และ IRI ซึ่งเป็นบริษัทผู้ให้บริการข้อมูลในการดำเนินธุรกิจต่างๆ ดังนั้นเมื่อคลังข้อมูลมีการใช้ข้อมูลทั้งจากระบบการดำเนินงานที่มีจำนวนหลายระบบและจากแหล่งข้อมูลภายนอกจะทำให้เราจะต้องทำการรวบรวมข้อมูลเหล่านั้นแล้วจัดเก็บไว้ในคลังข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อการใช้งานต่อไป

Data inconsistencies are removed; data from diverse operational applications is integrated.



รูปที่ 2-2 การรวมกันของข้อมูล

รูปที่ 2-2 แสดงถึงกระบวนการรวบรวมข้อมูลอย่างง่ายของบริษัท/สถาบันทางการเงิน โดยข้อมูลนั้นจะถูกรวบรวมจาก 3 แอปพลิเคชัน แล้วเก็บไว้ในหัวข้อทางธุรกิจเกี่ยวกับบัญชี โดยแอปพลิเคชันที่เป็นแหล่งข้อมูลอินพุตของคลังข้อมูลจะประกอบไปด้วย



ซึ่งจากข้างต้นระบบทั้งสามอาจมีการตั้งชื่อต่างๆ ให้กับฟิลด์หรือแอททริบิวของข้อมูลที่แตกต่างกัน หรืออาจมีรูปแบบ (format) ของข้อมูลที่แตกต่างกัน อาทิ หมายเลขบัญชีสำหรับระบบ saving account ควรจะมีด้วยกัน 8 หลักด้วยกัน แต่สำหรับระบบ checking account จะใช้หมายเลขบัญชีเพียง 6 หลัก ซึ่งจากความแตกต่างข้างต้น เราจะต้องกำจัดความไม่สอดคล้องกันของข้อมูลโดยการสร้างมาตรฐานให้กับข้อมูลเหล่านั้นเพื่อให้ข้อมูลที่เหมือนกันที่ถูกเก็บโดยแอปพลิเคชันที่แตกต่างกันและถูกจัดเก็บลงในฐานข้อมูลที่แตกต่างกันให้สื่อความหมายเดียวกัน ดังนั้นก่อนที่จะทำการเคลื่อนย้ายข้อมูลเข้าสู่คลังข้อมูลเราจะต้องทำการเปลี่ยนแปลง/เปลี่ยนรูปข้อมูลให้เป็นมาตรฐานเสียก่อน

3

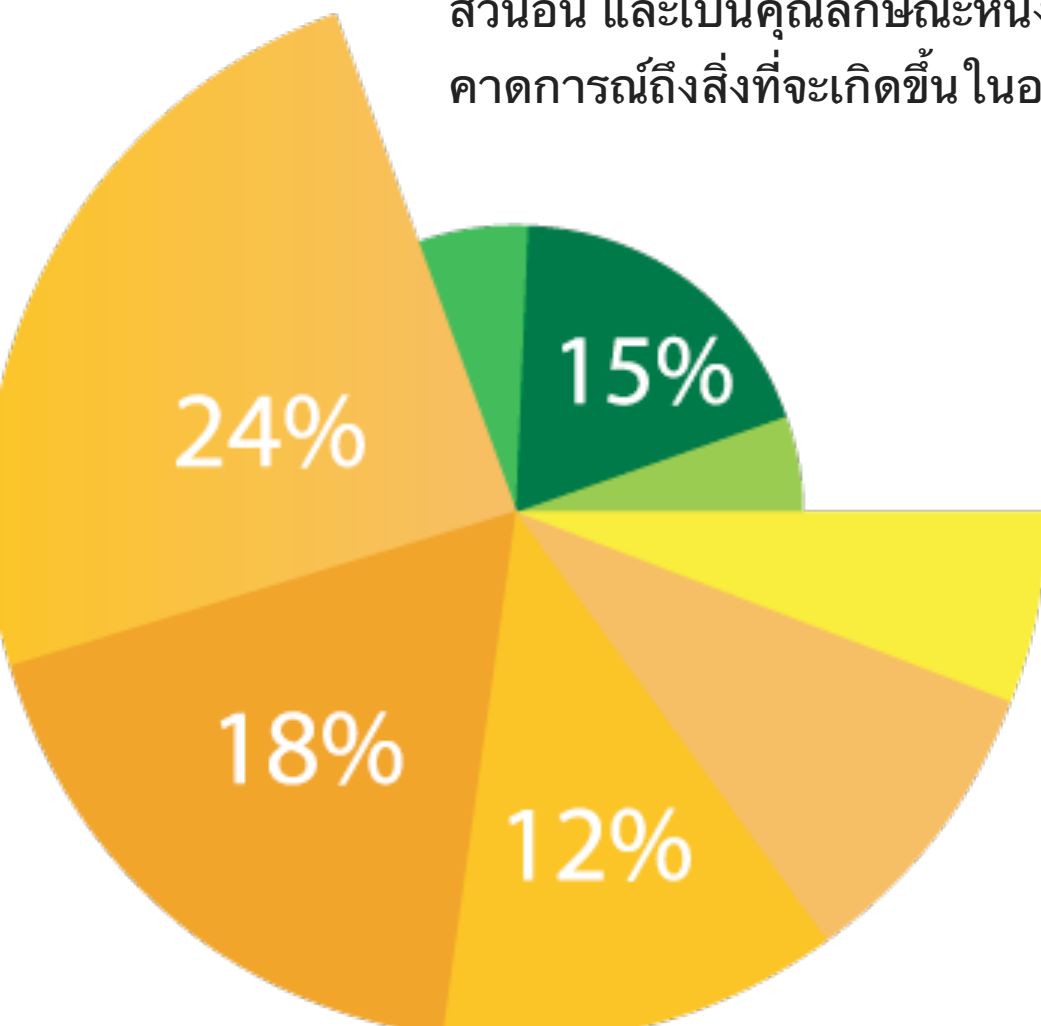
ข้อมูลที่เกี่ยวข้องกับช่วงเวลาต่างๆ (Time-variant data)

การจัดเก็บข้อมูลของระบบการดำเนินงานจะทำการเก็บข้อมูลที่เป็นปัจจุบันเท่านั้น เช่น ระบบบัญชีของธนาคาร ที่ทำการเก็บข้อมูลยอดเงินคงเหลือปัจจุบันของบัญชีลูกค้า ระบบการส่งสินค้าจะทำการเก็บข้อมูลการส่งสินค้าครั้งล่าสุดของลูกค้ารายหนึ่งๆ เป็นต้น

แต่อย่างไรก็ตามระบบการดำเนินงานบางระบบอาจจะทำการเก็บข้อมูลในอดีตบ้างแต่ก็เป็นการเก็บข้อมูลเพื่อสนับสนุนการทำธุรกิจในแต่ละวันเท่านั้น

แต่ในส่วน of ข้อมูลในคลังข้อมูล ที่ทำการสร้าง/จัดเตรียมข้อมูลเชิงกลยุทธ์เพื่อช่วยเหลือผู้ใช้ในการวิเคราะห์ข้อมูลในหลายๆ แง่มุม เช่น ผู้ใช้คลังข้อมูลอาจจะต้องการรูปแบบการซื้อสินค้า (buying pattern) ของลูกค้าแต่ละราย โดยข้อมูลที่ผู้ใช้สนใจจะไม่ได้เป็นเพียงแค่ข้อมูลการซื้อสินค้าครั้งล่าสุดของลูกค้าแต่ละรายเท่านั้น แต่จะสนใจข้อมูลการซื้อครั้งก่อนๆ หน้าด้วย หรือ ในอีกกรณีหนึ่งที่ผู้ใช้คลังข้อมูลอาจต้องการทราบถึงเหตุผลที่ยอดขายสินค้าลดลงในแถบตะวันออกเฉียงเหนือ ด้วยความต้องการดังกล่าวผู้ใช้จะต้องการข้อมูลยอดขายทั้งหมดที่เกิดขึ้นในเขตตะวันออกเฉียงเหนือที่เกิดขึ้นในช่วงเวลาที่ผ่านมา และในอีกกรณีหนึ่งนักการตลาดของบริษัทที่ทำธุรกิจร้านค้าปลีกจะต้องการที่จะโปรโมทสินค้า 2 รายการหรือมากกว่านั้น นักวิเคราะห์อาจต้องการยอดขายของรายการสินค้าที่ต้องการโปรโมทเทียบกับรายการสินค้าอื่นๆ ในแต่ละช่วงไตรมาสที่ผ่านมา เป็นต้น ซึ่งจากความต้องการที่ค่อนข้างจะหลากหลายของผู้ใช้ คลังข้อมูลจะต้องทำการเก็บข้อมูลที่เป็นปัจจุบัน และข้อมูลย้อนหลัง โดยมีข้อมูลแกนเวลาเข้ามาเกี่ยวข้อง ซึ่งการเก็บข้อมูลลักษณะนี้จะช่วยให้ผู้ใช้สามารถทราบถึงความเปลี่ยนแปลงของข้อมูลต่อช่วงเวลาต่างๆ ได้

ตัวอย่างเช่น ในคลังข้อมูลที่ประกอบไปด้วยข้อมูลยอดขายที่เป็นจำนวนชิ้นสินค้า ซึ่งในการจัดเก็บข้อมูลการขายสินค้าอาจจะมีการเก็บเวลาที่เกี่ยวข้องกับข้อมูลการขายสินค้านั้นๆ แนบไปกับข้อมูลจริงที่ต้องทำการเก็บ โดยเวลาที่ถูกจัดเก็บจะมีความละเอียดที่หลากหลาย เช่น ช่วงเวลา วัน เดือน ปี ที่มีการขายสินค้า เป็นต้น โดยในการจัดเก็บข้อมูลเราอาจจำเป็นต้องทำการจัดเก็บข้อมูลที่เป็นจำนวนชิ้นสินค้าที่ขายได้กับหน่วยของเวลา เพื่อบ่งบอกถึงยอดขายรายวัน ยอดขายแต่ละสัปดาห์ ยอดขายแต่ละเดือน หรือยอดขายแต่ละไตรมาส เป็นต้น ซึ่งจากตัวอย่างข้างต้นเราจะเห็นว่า **“time-variant”** นั้นจะเป็นจะเป็นการจัดเก็บข้อมูลเชิงเวลาแนบไว้กับข้อมูลส่วนอื่น และเป็นคุณลักษณะหนึ่งของคลังข้อมูลที่ทำให้ผู้ใช้สามารถวิเคราะห์ข้อมูลในอดีตและปัจจุบันเพื่อคาดการณ์ถึงสิ่งที่จะเกิดขึ้นในอนาคตได้



4

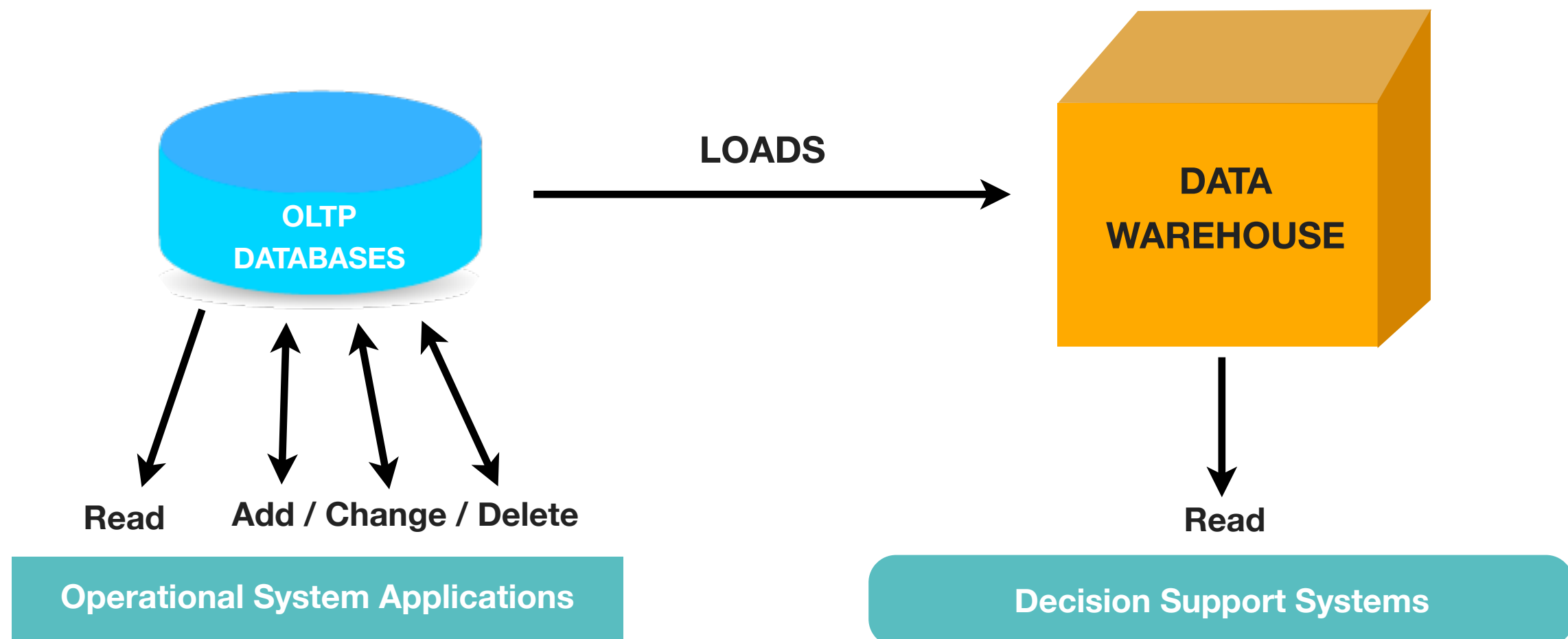
ข้อมูลที่ไม่เปลี่ยนแปลง
(Nonvolatile data)

อย่างที่เรทราบจากบทที่แล้วว่า ระบบการดำเนินงานจะอนุญาตให้ผู้ใช้สามารถทำการเรียกดูข้อมูล (Select) เพิ่มข้อมูลลงในฐานข้อมูล (Insert) ลบข้อมูลลงในฐานข้อมูล (Delete) และอัปเดตข้อมูลต่างๆ (Update) ได้ ซึ่งเราจะสามารถดำเนินการได้ทุกๆ การทำงาน แต่ในส่วน of คลังข้อมูล การทำงานจะเริ่มจากการเลือกหรือสกัดข้อมูลที่ต้องการเพียงบางส่วนจากระบบการดำเนินงานและแหล่งข้อมูลอื่นๆ จากนั้นทำการรวบรวมข้อมูลเข้าด้วยกันและทำการประมวลผลข้อมูล จากนั้นค่อยทำการจัดเก็บข้อมูลลงในฐานข้อมูลเพื่อให้ผู้ใช้สามารถเรียกใช้ข้อมูลสำหรับการวิเคราะห์ต่างๆ ได้

Select Insert
Delete
Insert Update
Select

จากขั้นตอนการทำงานของคลังข้อมูลดังกล่าว ข้อมูลจากระบบการดำเนินงานจะถูกเลือกออกมาเพียงบางส่วนและถูกเคลื่อนย้ายเข้าสู่คลังข้อมูลในช่วงเวลาที่กำหนด โดยขึ้นอยู่กับความต้องการทางธุรกิจ เช่น ทำการเคลื่อนย้ายข้อมูลวันละสองครั้ง วันละครั้ง อาทิตย์ละครั้ง หรือสองอาทิตย์หนึ่งครั้ง เป็นต้น ซึ่งจากการทำงานดังกล่าว จะทำให้เห็นภาพได้ว่าขั้นตอนการทำงานแทบทั้งหมดจะถูกดำเนินการในระบบ ผู้ใช้งานคลังข้อมูลจะไม่สามารถทำการเพิ่ม ลบ หรืออัปเดตข้อมูลในคลังข้อมูลได้ ผู้ใช้งานจะสามารถเรียกดูข้อมูลได้เท่านั้น เพื่อให้เข้าใจมากขึ้น ลองพิจารณารูปที่ 2-3 ที่การทำธุรกรรมทางธุรกิจในแต่ละครั้งจะมีการอัปเดตระบบการดำเนินงานแบบทันที และผู้ใช้ระบบการดำเนินงานสามารถเพิ่ม เปลี่ยนแปลง และลบข้อมูลออกจากระบบการดำเนินงานได้ แต่สำหรับคลังข้อมูล จะไม่ทำการอัปเดตข้อมูลแบบทันที แต่จะทำการอัปเดตตามเวลาที่กำหนด และผู้ใช้ไม่สามารถลบข้อมูลออกจากคลังข้อมูลได้

Data inconsistencies are removed; data from diverse operational applications is integrated.



รูปที่ 2-3 คุณสมบัติการไม่เปลี่ยนแปลงข้อมูลในคลังข้อมูล

5

ข้อมูลที่มีรายละเอียดหลายระดับ (Data granularity)

การเก็บข้อมูลในระบบการดำเนินงานมักจะทำการเก็บข้อมูลในลักษณะที่มีความละเอียดค่อนข้างสูง อาทิเช่น การขายของร้านค้าปลีกจะทำการเก็บจำนวนสินค้าที่ขายได้ในแต่ละรายการที่จุดแคชเชียร์คิดเงิน หรือการสั่งสินค้า จะทำการเก็บจำนวนสินค้าที่สั่งในแต่ละครั้ง เป็นต้น โดยส่วนใหญ่ของระบบการดำเนินงานจะไม่ทำการเก็บข้อมูลที่เป็นผลสรุปแต่จะเน้นที่การเก็บข้อมูลแต่ละรายการ (transaction) เพื่อการดำเนินธุรกิจในแต่ละวัน

แต่เมื่อไรก็ตามที่เราต้องการข้อมูลที่เป็นผลสรุปจากระบบการดำเนินงาน เราจะต้องทำการรวมข้อมูลแต่ละรายการเข้าด้วยกัน เช่น ถ้าเราต้องการยอดขายของสินค้าชนิดหนึ่งในเดือนมิถุนายน เราจะต้องทำการอ่านข้อมูลการขายสินค้าทั้งหมดของเดือนมิถุนายนจากนั้นทำการรวมข้อมูลการขายสินค้าเหล่านั้นเพื่อให้ได้เป็นข้อมูลที่เป็นผลสรุปที่เราต้องการ แต่ในกรณีของคลังข้อมูลที่ใช้ผู้ใช้งานมักจะทำการเรียกดูข้อมูลที่เป็นผลสรุป ซึ่งผู้ใช้งานอาจทำการเรียกดูข้อมูลจำนวนสินค้ารายการหนึ่งๆ ที่ขายได้ในภาคตะวันออก จากนั้นค่อยเพิ่มความละเอียดของข้อมูลที่ต้องการได้รับ ขึ้น ซึ่งข้อมูลที่ต้องการอาจเป็นยอดขายของสินค้ารายการหนึ่งที่ได้ขายได้ในแต่ละจังหวัดของภาคตะวันออก และอาจเจาะลึกไปถึงข้อมูลยอดขายของสินค้ารายการหนึ่งที่ขายได้ในแต่ละสาขาที่อยู่ในภาคตะวันออก เป็นต้น ซึ่งโดยส่วนใหญ่แล้วผู้ใช้งานมักจะเริ่มจากการเรียกดูข้อมูลที่มีรายละเอียดน้อยแล้วค่อยๆ เพิ่มรายละเอียดขึ้นเรื่อย ๆ



จากการใช้งานคลังข้อมูลข้างต้น การจัดเก็บข้อมูลในคลังข้อมูลโดยส่วนใหญ่จะทำการเก็บข้อมูลที่เป็นแบบผลรวมหรือผลสรุปที่มีความละเอียดแตกต่างกันตามความต้องการของผู้ใช้ ซึ่งถ้าคลังข้อมูลมีการจัดเก็บข้อมูลในลักษณะที่มีความละเอียดสูงจะทำให้เราต้องทำการเก็บข้อมูลเป็นจำนวนมาก ดังนั้นในการจัดเก็บข้อมูล เราจะต้องทำการกำหนดระดับความละเอียดของข้อมูลตามชนิดข้อมูลในคลังข้อมูล และทำการพิจารณาถึงประสิทธิภาพที่คาดหวังจากการค้นคืนผลลัพธ์จากคลังข้อมูล

ในการที่จะทำความเข้าใจเกี่ยวกับระดับความละเอียดของข้อมูลในคลังข้อมูล ลองพิจารณารูปที่ 2-4 ที่ประกอบไปด้วยการจัดเก็บข้อมูลในคลังข้อมูลของธนาคารหนึ่งๆ ที่มีความละเอียดแตกต่างกัน 3 ระดับ นั่นคือ การเก็บข้อมูลรายวัน รายเดือน และรายไตรมาส ตามลำดับ

THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

<u>Daily Detail</u>	<u>Monthly Summary</u>	<u>Quarterly Summary</u>
Account	Account	Account
Activity Date	Month	Quarter
Amount	Number of transactions	Number of transactions
Deposit/Withdrawal	Withdrawals	Withdrawals
	Deposits	Deposits
	Beginning Balance	Beginning Balance
	Ending Balance	Ending Balance

Data granularity refers to the level of detail. Depending on the requirements, multiple levels of detail may be present.

Many data warehouses have at least dual levels of granularity.

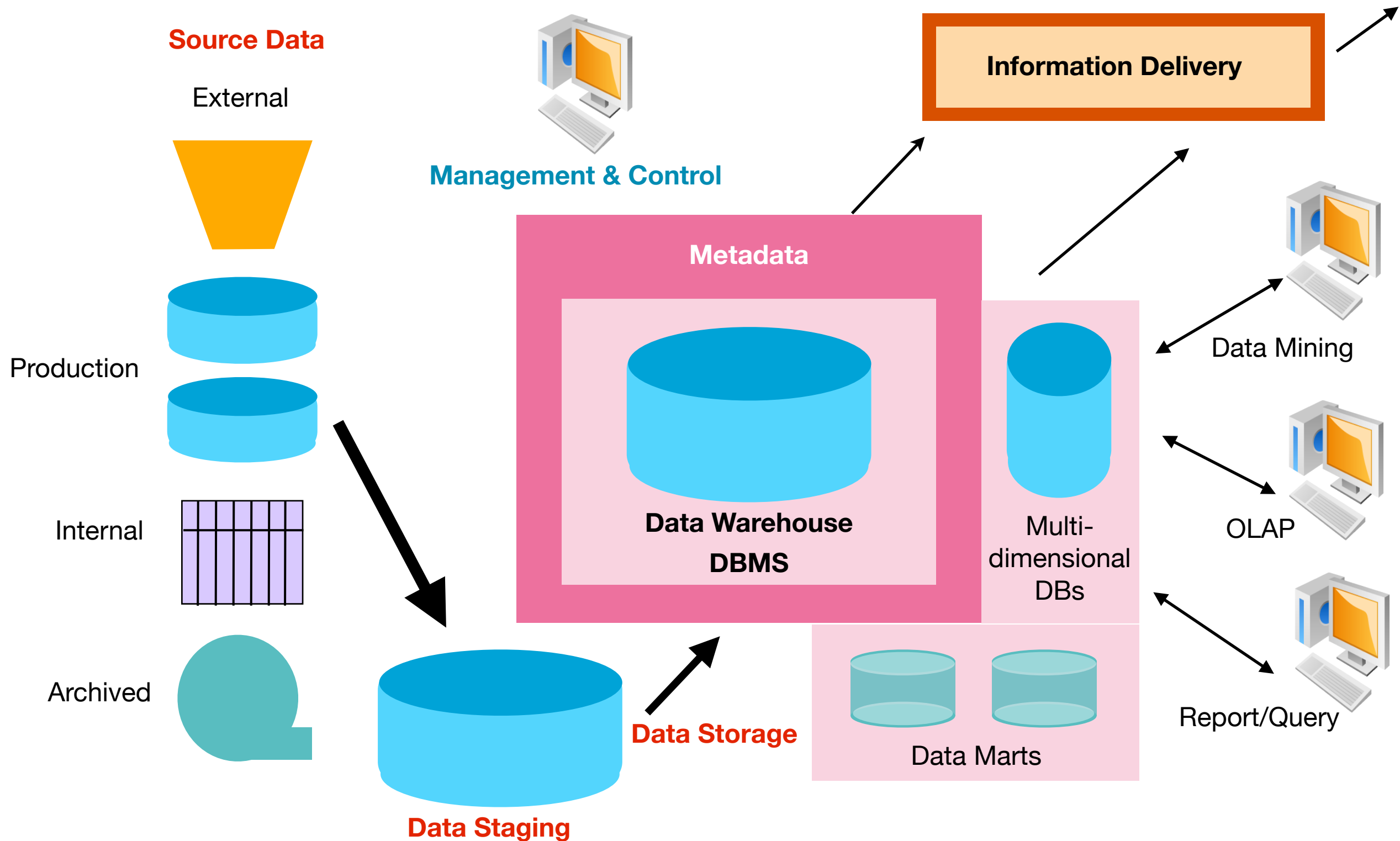
รูปที่ 2-4 ความละเอียดของข้อมูลในคลังข้อมูล

ส่วนประกอบของคลังข้อมูล

หลังจากทราบถึงคุณลักษณะต่างๆ รวมถึงฟังก์ชันการทำงานหลักของคลังข้อมูล ในส่วนนี้จะพิจารณาถึงส่วนประกอบต่างๆของคลังข้อมูล ซึ่งในการสร้างคลังข้อมูลจะคิด พิจารณา และเกี่ยวข้องกับส่วนประกอบต่างๆ ทั้งในส่วนของการซอร์ฟแวร์และฮาร์ดแวร์ โดยเราจะต้องทำการรวมส่วนประกอบเหล่านี้เข้าด้วยกันและทำการปรับแต่งส่วนประกอบเหล่านี้เพื่อให้คลังข้อมูลมีการทำงานที่มีประสิทธิภาพและประโยชน์สูงสุด

การเลือกส่วนประกอบที่จะใช้ในคลังข้อมูลจะขึ้นอยู่กับข้อจำกัดและความต้องการของแต่ละองค์กรเป็นหลัก โดยส่วนประกอบพื้นฐานของคลังข้อมูลจะถูกแสดงในรูปที่ 2-5 ซึ่งจากทางซ้ายสุดคือแหล่งข้อมูล (Source data) และ ถัดมาคือ “data staging” หรือ “staging area” ที่เป็นตัวกลางหรือเป็นที่พักข้อมูลที่ถูกสกัด/ถูกเลือกมาจากแหล่งข้อมูล จากนั้นจะเป็นส่วนของพื้นที่ในการจัดเก็บข้อมูลและเมตาดาต้า (Data and metadata storage) และทางด้านขวาสุดจะเป็นส่วนของระบบที่ใช้สำหรับเข้าถึงหรือส่งผ่านข้อมูลไปยังผู้ใช้งาน (Information delivery) ที่ประกอบไปด้วยวิธีการต่างๆ สำหรับการส่งข้อมูลให้กับผู้ใช้ และยังรวมถึงเครื่องมือต่างๆที่ใช้ในการวิเคราะห์ข้อมูลที่ซับซ้อน อาทิ เครื่องมือสำหรับการทำเหมืองข้อมูล (Data mining) เครื่องมือสำหรับวิเคราะห์ข้อมูลแบบออนไลน์ (Online analytical processing, OLAP) และ เครื่องมือในการสร้างคิวรีและรายงานต่างๆ (Query and report tools) เป็นต้น

Architecture is the proper arrangement of the components.



ในการที่จะทำความเข้าใจส่วนประกอบหลักทั้ง 4 ส่วน ดังแสดงในรูปที่ 2-5 เราควรที่จะต้องทำการศึกษาถึงรายละเอียดของแต่ละส่วนที่จะใช้ในการสร้างคลังข้อมูล โดยรายละเอียดของแต่ละส่วนจะสามารถอธิบายได้ดังนี้

- แหล่งข้อมูลของคลังข้อมูล
- พื้นที่พักข้อมูล
- พื้นที่สำหรับจัดเก็บข้อมูล
- ระบบเข้าถึงและส่งผ่านข้อมูลไปยังผู้ใช้
- ส่วนงานการจัดการและการควบคุมต่างๆ
- การจัดเก็บเมตาดาต้า

● แหล่งข้อมูลของคลังข้อมูล

แหล่งข้อมูลของคลังข้อมูลสามารถแบ่งได้เป็น 4 ประเภท ดังนี้

Production data

Internal data

Archieved data

External data



Production data จะเป็นข้อมูลที่มาจากระบบการดำเนินงานหลายระบบด้วยกัน เช่น ระบบการเงิน ระบบการผลิต ระบบการสั่งซื้อสินค้า ระบบตลาดห่วงโซ่อุปทาน และระบบการจัดการความสัมพันธ์ลูกค้า เป็นต้น

ข้อมูลเหล่านี้จะถูกเลือกหรือสกัดมาจากระบบการดำเนินงานโดยทำการเลือกจากความต้องการข้อมูลในคลังข้อมูล แต่ในการเลือกข้อมูลจากระบบการดำเนินงานหลายๆ ระบบอาจทำให้เราต้องพบกับข้อมูลที่มีรูปแบบที่หลากหลาย เช่น ข้อมูลอาจมาจากฮาร์ดแวร์ที่แตกต่างกัน ระบบฐานข้อมูลที่แตกต่างกัน ระบบปฏิบัติการที่ต่างกัน เป็นต้น โดยข้อมูลที่ได้รับจากระบบการดำเนินงานที่แตกต่างกันอาจมีความไม่สอดคล้องกันของข้อมูลเจือปนอยู่ด้วย ดังนั้นเมื่อเราได้รับข้อมูลจากระบบการดำเนินงานแล้วเราจะต้องทำให้ข้อมูลเหล่านั้นเป็นมาตรฐานเดียวกันด้วย

● แหล่งข้อมูลของคลังข้อมูล

แหล่งข้อมูลของคลังข้อมูลสามารถแบ่งได้เป็น 4 ประเภท ดังนี้

Production data

Internal data

Archieved data

External data

Internal data จะเป็นข้อมูลประเภทสเปรดชีต เอกสารต่างๆ ที่แสดงรายละเอียดของลูกค้าหรือ ฐานข้อมูลของแผนกที่ถูกสร้างขึ้นไว้ใช้งานส่วนตัว ซึ่งถูกเก็บไว้ใช้ในการดำเนินงานบางอย่างของการดำเนินธุรกิจในแต่ละแผนก ข้อมูลเหล่านี้จะเป็นข้อมูลที่อาจมีความสำคัญที่เราไม่สามารถละเลยได้ ในการเก็บรวบรวมข้อมูลที่เป็น “**internal data**” จากผู้ใช้ เราอาจต้องทำการตัดสินใจว่าเราควรเก็บ internal data เป็นจำนวนเท่าใด โดยข้อมูลเหล่านี้จะทำให้ข้อมูลในคลังข้อมูลมีปริมาณเพิ่มขึ้นและยังเป็นการเพิ่มความซับซ้อนให้กับขั้นตอนการทำงานของคลังข้อมูลอีกด้วย (จะทำให้ขั้นตอนการรวบรวมข้อมูล และขั้นตอนการทำให้ข้อมูลเป็นมาตรฐานจะมีความซับซ้อนและยุ่งยากมากขึ้น)

ดังนั้น ในการที่จะพิจารณาที่จะจัดเก็บข้อมูลที่เป็น “**internal data**” เราควรที่จะต้องหาหรือออกแบบวิธีการเลือกข้อมูลเหล่านี้ ค้นหาวิธีในการเข้าถึงและสกัดข้อมูลจากเอกสารต่างๆ และ พิจารณาการรวมฐานข้อมูลย่อยๆ ที่ถูกจัดเก็บไว้ในแต่ละแผนกเข้าด้วยกัน



● แหล่งข้อมูลของคลังข้อมูล

แหล่งข้อมูลของคลังข้อมูลสามารถแบ่งได้เป็น 4 ประเภท ดังนี้

Production data

Internal data

Archieved data

External data

Archieved data จะเป็นข้อมูลเก่าที่อาจมีอายุมากกว่า 1-5 ปีขึ้นไป โดยข้อมูลเหล่านี้อาจจะไม่ได้ถูกจัดเก็บอยู่ในฐานข้อมูลที่จะมีข้อมูลใหม่ๆ แต่จะถูกจัดเก็บไว้ในแฟ้มข้อมูล ดิสก์ หรือเทป เป็นต้น

ข้อมูลที่เป็น “**archieved data**” จะเป็นข้อมูลชนิดหนึ่งที่มีความสำคัญกับคลังข้อมูล เนื่องจากคลังข้อมูลจะมีคุณลักษณะหนึ่งที่มีการพิจารณาข้อมูลย้อนหลังเพื่อใช้ในการวิเคราะห์รูปแบบของข้อมูลและวิเคราะห์แนวโน้มของข้อมูล ดังนั้น เราอาจจำเป็นต้องทำการเก็บข้อมูลที่ค่อนข้างเก่าเหล่านี้ไว้ในคลังข้อมูลด้วยเช่นกัน



● แหล่งข้อมูลของคลังข้อมูล

แหล่งข้อมูลของคลังข้อมูลสามารถแบ่งได้เป็น 4 ประเภท ดังนี้

Production data

Internal data

Archieved data

External data



External data จะเป็นข้อมูลเกี่ยวกับสถิติในภาคอุตสาหกรรมที่สร้างขึ้นจากบริษัทภายนอก และ หน่วยงานราชการต่างๆ ที่ผู้บริหารจะใช้ข้อมูลเหล่านี้เพื่อใช้ประกอบการตัดสินใจบางอย่าง ด้วยเหตุนี้ในการสร้างคลังข้อมูลเราอาจจำเป็นต้องทำการเก็บข้อมูลจากแหล่งข้อมูลภายนอกไว้ เพื่อทำการวิเคราะห์แนวโน้มของอุตสาหกรรม และเปรียบเทียบประสิทธิภาพระหว่างบริษัทของเรากับองค์กรอื่นๆ เป็นต้น

แต่อย่างไรก็ดีข้อมูลจากภายนอกจะมีการจัดรูปแบบของข้อมูลที่ไม่เหมือนกับข้อมูลภายใน เราจะต้องทำการออกแบบวิธีในการแปลงข้อมูลให้อยู่ในรูปแบบเดียวกับข้อมูลภายใน และเราอาจจำเป็นต้องจัดการกับการส่งผ่านข้อมูลจากแหล่งข้อมูลภายนอกอีกด้วย

● พื้นที่พักข้อมูล

หลังจากที่เราทำการเลือกหรือสกัดข้อมูลที่ต้องการจากระบบการดำเนินงานต่างๆ และจากแหล่งข้อมูลภายนอกแล้ว เราจะต้องทำการประมวลผลข้อมูลเหล่านั้นก่อนที่จะทำการนำข้อมูลเหล่านั้นไปจัดเก็บไว้ในคลังข้อมูล โดยในการประมวลผลข้อมูลเราจะต้องทำการเปลี่ยน แปลง และจัดข้อมูลให้อยู่ในรูปแบบเหมาะสมต่อโครงสร้างข้อมูลคลังข้อมูล แต่เนื่องจากข้อมูลในระบบการดำเนินงานและคลังข้อมูลมีความแตกต่างกัน และถูกแยกออกจากกัน

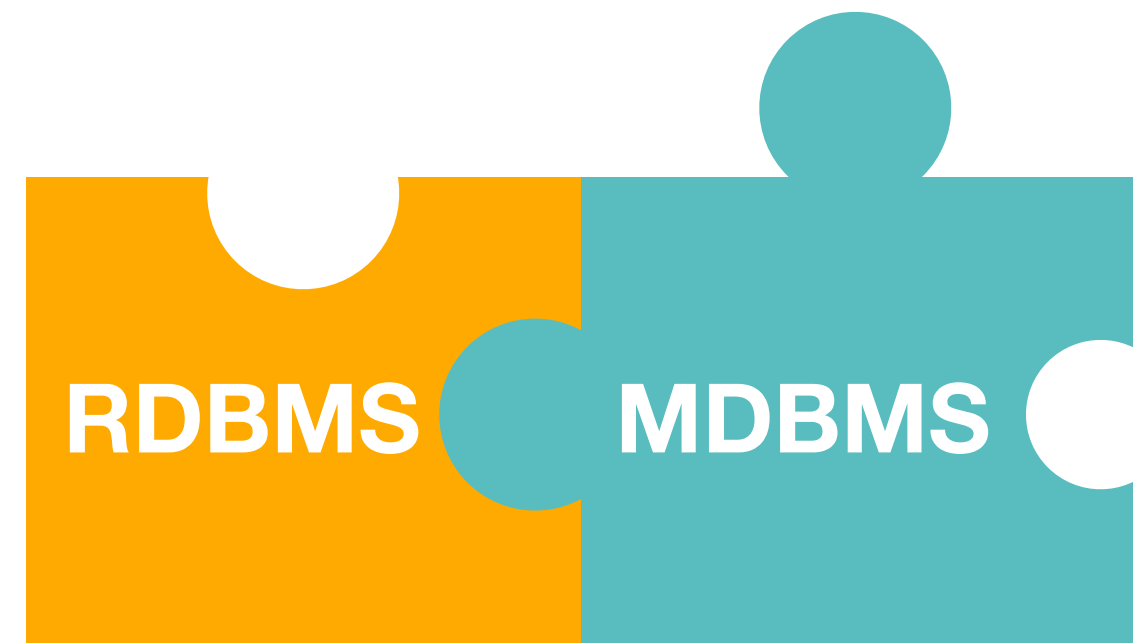
กล่าวคือ ข้อมูลในระบบการดำเนินงานจะถูกจัดเก็บตามแอปพลิเคชันเพื่อสนับสนุนการดำเนินธุรกิจในแต่ละวัน แต่ข้อมูลในคลังข้อมูลจะถูกจัดเก็บตามหัวข้อทางธุรกิจ ซึ่งข้อมูลอาจมาจากการสรุปข้อมูลในแง่มุมต่างๆ เมื่อข้อมูลของทั้งสองระบบมีความแตกต่างกัน เราจึงจำเป็นต้องสร้างที่พักข้อมูล ที่อยู่ระหว่างระบบการดำเนินงานและคลังข้อมูล ซึ่งก็คือ “data staging” หรือ “staging area” ที่จะประกอบไปด้วยพื้นที่สำหรับจัดเก็บข้อมูลที่สกัดได้จากระบบการดำเนินงาน และฟังก์ชันการทำงานต่างๆ เช่น การทำความสะอาดข้อมูล การเปลี่ยนแปลงข้อมูล การรวมข้อมูลเข้าด้วยกัน เพื่อทำการเตรียมข้อมูลสำหรับจัดเก็บในคลังข้อมูลต่อไป



● พื้นที่สำหรับจัดเก็บข้อมูล

ในส่วนของพื้นที่สำหรับจัดเก็บข้อมูลในคลังข้อมูลจะเป็นส่วนที่แยกออกมาจากระบบการดำเนินงาน โดยในการจัดเก็บข้อมูล เราอาจเรียกใช้เครื่องมือต่างๆ ที่มีวางจำหน่าย หรือทำการสร้างฟังก์ชันการจัดเก็บข้อมูลขึ้นเอง ซึ่งโดยส่วนใหญ่แล้วคลังข้อมูลจะใช้ “**Relational DBMS, RDBMS**” ในการจัดเก็บข้อมูล แต่ก็มีบางคลังข้อมูลใช้ “**Multidimensional DBMS, MDBMS**” เพื่อเก็บข้อมูลด้วยเช่นกัน

ในการจัดเก็บข้อมูล ผู้สร้างคลังข้อมูลควรจะเน้นย้ำที่การจัดเก็บข้อมูลทั้งในปัจจุบันและข้อมูลย้อนหลัง รวมถึงการวางแผนหรือออกแบบโครงสร้างของข้อมูลที่จะใช้ในการวิเคราะห์ข้อมูลเหล่านั้น นอกจากนี้ยังต้องคำนึงถึงประสิทธิภาพในการเรียกใช้ข้อมูลอีกด้วย



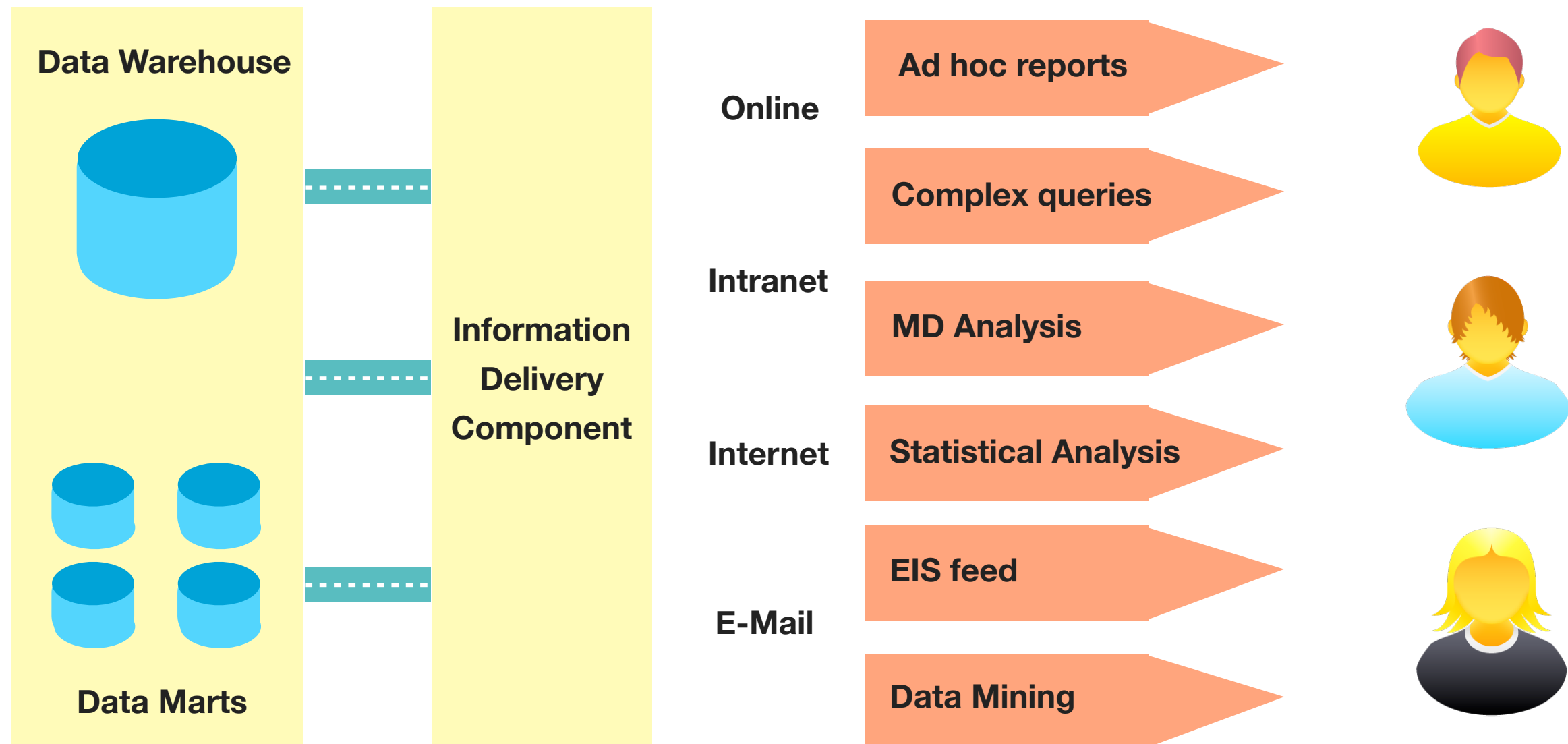
● ระบบเข้าถึงและส่งผ่านข้อมูลไปยังผู้ใช้

ในการพิจารณาเกี่ยวกับการส่งผ่านข้อมูลให้กับผู้ใช้ เราจะต้องทราบก่อนว่าคลังข้อมูลที่เราสร้างขึ้นนั้นมีผู้ใช้กี่ประเภท? แต่ละประเภทเป็นใครบ้าง? และผู้ใช้แต่ละประเภทหรือแต่ละรายต้องการข้อมูลประเภทใด? โดยส่วนใหญ่ของผู้ใช้ที่เพิ่งเริ่มใช้งานที่ยังไม่ได้ผ่านการอบรมและยังไม่มีประสบการณ์ในการใช้งานคลังข้อมูลมาก่อน รวมถึงผู้ใช้งานคลังข้อมูลแบบเป็นครั้งคราวมักจะต้องการรายงานและการประมวลผลคิวรีที่ระบบคลังข้อมูลกำหนดไว้ให้หรือจัดเตรียมไว้ให้อยู่แล้ว ในขณะที่นักวิเคราะห์ทางธุรกิจต้องการที่จะวิเคราะห์ข้อมูลที่มีความซับซ้อน และผู้ใช้ที่มีอำนาจในการตัดสินใจ มักจะต้องการที่จะเรียกดูข้อมูลที่น่าสนใจ เป็นต้น

เมื่อผู้ใช้งานมีหลายกลุ่มและมีความต้องการที่หลากหลาย เราอาจจำเป็นต้องทำการออกแบบหรือกำหนดฟังก์ชันการส่งข้อมูลที่แตกต่างกันเพื่อตอบสนองต่อการส่งข้อมูลที่หลากหลายให้กับผู้ใช้แต่ละประเภท ดังแสดงในรูปที่ 2-6 ที่แสดงฟังก์ชันการส่งข้อมูล 4 วิธีด้วยกัน

ซึ่ง โดยส่วนใหญ่การส่งข้อมูลที่เป็นคิวรีและรายงานต่างๆ จะเป็นแบบออนไลน์ที่อนุญาตให้ผู้ใช้สามารถรับข้อมูลได้อย่างทันทีที่มีการส่งคิวรีที่ต้องการไปประมวลผลที่คลังข้อมูล อีกวิธีการหนึ่งที่ได้รับคามนิยมลดหลั่นลงมาคือการตั้งเวลาในการส่งคิวรีไปยังคลังข้อมูลเพื่อประมวลผล และการเรียกดูรายงานตามช่วงเวลาที่กำหนด โดยหลังจากทำการตั้งเวลาแล้วผู้ใช้จะได้รับรายงานต่างๆ ผ่านทางอีเมล หรือเราอาจใช้อินเทอร์เน็ตของบริษัทในการเรียกดูข้อมูลได้เช่นกัน ซึ่งในปัจจุบันเทคโนโลยี อินทราเน็ตและอินเทอร์เน็ตจะเป็นวิธีที่ได้รับความนิยมอย่างมากในองค์กรใหญ่ๆ





รูปที่ 2-6 ระบบการเข้าถึงและส่งผ่านข้อมูลไปยังผู้ใช้งาน

● ส่วนงานการจัดการและการควบคุมต่าง ๆ

ส่วนงานการจัดการและควบคุมจะทำหน้าที่ประสานงานในกิจกรรมและการให้บริการต่างๆ ภายในคลังข้อมูล การทำงานของส่วนงานนี้จะประกอบไปด้วยการควบคุมการเปลี่ยนแปลง/เปลี่ยนรูปข้อมูล การถ่ายโอนข้อมูลไปยังคลังข้อมูล และการควบคุมการส่งข้อมูลให้กับผู้ใช้ นอกจากนี้ยังเป็นการประกอบการทำงานร่วมกับระบบจัดการฐานข้อมูล (DBMS) จัดการให้ข้อมูลถูกจัดเก็บในที่เก็บข้อมูลอย่างเหมาะสม คอยเฝ้าดูการเคลื่อนที่ของข้อมูลไปยัง staging area และข้อมูลที่ออกจาก staging Area ไปยังที่เก็บข้อมูลของคลังข้อมูลอีกด้วย



● การจัดเก็บเมตาดาต้า

เมตาดาต้าในคลังข้อมูลเปรียบเสมือนพจนานุกรมข้อมูลหรือแคตตาล็อกของข้อมูลในระบบจัดการฐานข้อมูล (Data dictionary/data catalog of DBMS) ซึ่งภายในพจนานุกรมข้อมูลจะมีการจัดเก็บข้อมูลที่เกี่ยวข้องกับ โครงสร้างการจัดเก็บข้อมูลต่างๆ เช่น การจัดเก็บข้อมูลเกี่ยวกับแฟ้มที่ทำการเก็บข้อมูล รวมถึง address ต่างๆ ข้อมูลเกี่ยวกับ index และ อื่นๆ

นอกจากนั้น เรายังสามารถเปรียบเทียบเมตาดาต้าในเชิงกว้างๆ ได้ ซึ่งเมตาดาตานั้นสามารถเปรียบได้กับข้อมูลตำแหน่งของเมืองที่เราอยู่ซึ่งในบางครั้งอาจต้องการข้อมูลเกี่ยวกับห้างร้านต่างๆ ในเมืองของคุณว่าร้านแห่งนั้นอยู่ที่ใด ชื่ออะไรบ้าง มีสินค้าอะไรในร้านเหล่านั้นบ้าง เมื่อเราต้องการข้อมูลเราควรเปิดสมุดหน้าเหลือง เมตาดาต้าทำหน้าที่เหมือนกับไดเรกทอรีของข้อมูลสำหรับคลังข้อมูลของคุณ ดังนั้น เมตาดาต้าจึงมีความสำคัญสำหรับการสร้างและการใช้คลังข้อมูลเป็นอย่างมาก เช่น

- (1) เมตาดาต้าทำหน้าที่เสมือนการเชื่อมส่วนต่างๆ ของคลังข้อมูลเข้าด้วยกัน
- (2) เมตาดาต้าจะช่วยให้ผู้พัฒนาคลังข้อมูลเข้าใจถึงเนื้อหา/ข้อมูล และโครงสร้างของฐานข้อมูล
- (3) เมตาดาต้าจะช่วยให้ผู้ใช้สามารถจำเนื้อหา/ข้อมูลเฉพาะทาง/คำศัพท์ของพวกเขาได้ เป็นต้น



ข้อมูลที่เป็นเมตาดาต้าจะสามารถแบ่งกลุ่มได้เป็น 3 ชนิดหลักๆ ดังนี้

1

เมตาดาต้าที่ได้มาจาก
ระบบดำเนินงาน
(Operational metadata)

2

เมตาดาต้าที่ได้มาจากการ
เลือกข้อมูล และการ
เปลี่ยนแปลง/เปลี่ยนรูปข้อมูล
(Extract and transformation
metadata)

3

เมตาดาต้าสำหรับผู้ใช้ใน
การใช้งานคลังข้อมูล
(End-user metadata)

M A T A D A T A

1

เมตาดาต้าที่ได้มาจาก
ระบบดำเนินงาน
(Operational metadata)

เมตาดาต้าที่ได้มาจากระบบดำเนินงาน (Operational metadata) จะเป็นข้อมูลรายละเอียดต่างๆ ที่เกี่ยวข้องกับระบบการดำเนินงาน เช่น ชื่อฐานข้อมูลของระบบ ชื่อตารางต่างๆ รวมถึงชื่อฟิลด์หรือแอททริบิวต์ที่เราสนใจ โครงสร้างข้อมูล ชนิดของข้อมูล ความยาวของข้อมูลในแต่ละฟิลด์ที่เราสนใจ และอื่นๆ

2

เมตาดาต้าที่ได้มากจากการ
เลือกข้อมูล และการ
เปลี่ยนแปลง/เปลี่ยนรูปข้อมูล
(Extract and transformation
metadata)

เมตาดาต้าที่ได้มากจากการเลือกข้อมูล และการเปลี่ยนแปลง/เปลี่ยนรูปข้อมูล (Extract and transformation metadata) จะเป็นข้อมูลที่เกี่ยวข้องกับการเลือกข้อมูลจากแหล่งข้อมูล เมตาดาต้าลักษณะนี้จะประกอบไปด้วย ชื่อต่างๆ ของข้อมูล/ตาราง หรืออื่นๆ ที่เกี่ยวข้องกับการเลือกข้อมูล ความถี่ในการสกัดข้อมูล วิธีการสกัดข้อมูล และกฎทางธุรกิจ (Business Rule) สำหรับการสกัดข้อมูล เป็นต้น

3

เมตาดาต้าสำหรับผู้ใช้ในการ
ใช้งานคลังข้อมูล
(End-user metadata)

เมตาดาต้าสำหรับผู้ใช้ในการใช้งานคลังข้อมูล (End-user metadata) จะเปรียบเสมือนแผนที่ของคลังข้อมูลที่ช่วยให้ผู้ใช้สามารถค้นหาข้อมูลและสารสนเทศจากคลังข้อมูลได้ โดย “End-User metadata” จะอนุญาตและยอมให้ผู้ใช้ทำการใช้คำศัพท์เฉพาะที่เกี่ยวข้องกับธุรกิจในการเรียกดูข้อมูลจากคลังข้อมูลได้อีกด้วย

เมตาดาต้าทั้ง 3 ชนิดข้างต้นจะมีประโยชน์และมีวัตถุประสงค์ของการจัดเก็บข้อมูลที่แตกต่างกัน โดยในการเรียกใช้ข้อมูลเมตาดาตานั้นสามารถทำได้หลายแง่มุม เช่น เมื่อผู้ใช้ต้องการใช้งานคลังข้อมูล ผู้ใช้จะสามารถเรียกดูข้อมูลเมตาดาต้าเพื่อที่จะทราบถึงชื่อตารางหรือชื่อฟิลด์ที่เก็บข้อมูลที่พวกเขาต้องการได้ หรือ ในส่วนของส่วนงานการจัดการและการควบคุมต่างๆ (ดังแสดงในรูปที่ 2-5) จะมีการเรียกใช้เมตาดาต้าเพื่อทำการจัดการและควบคุมการทำงานต่างๆ เนื่องจากเมตาดาต้าจะมีส่วนของข้อมูลเกี่ยวกับแหล่งข้อมูล พารามิเตอร์ต่างๆ ขั้นตอนการทำงานต่าง ชื่อของฐานข้อมูล และอื่นๆ โดยที่ข้อมูลเหล่านี้จะใช้ในการจัดการ และควบคุมกระบวนการทำงานของคลังข้อมูลทั้งหมด

จากที่กล่าวมาทั้งหมดข้างต้น เราจะเห็นภาพกว้างๆ ว่าเมตาดาตานั้นมีความสำคัญต่อการสร้างการทำงาน และการใช้งานคลังข้อมูล ดังนั้นเราจะทำการศึกษาถึงรายละเอียดของเมตาดาต้าอีกครั้งหนึ่งในบทที่ 9

คลังข้อมูลและดาต้ามาร์ท



หลังจากที่เราทราบถึงคุณลักษณะและส่วนประกอบต่างๆ ของคลังข้อมูลแล้ว เราจะสามารถกล่าวได้ว่า “คลังข้อมูลนั้นเป็นระบบสำหรับสร้างหรือจัดเตรียมข้อมูลที่เป็นผลสรุป” โดยข้อมูลที่ถูกจัดเก็บอยู่ในคลังข้อมูลจะถูกจัดเก็บตามหัวข้อทางธุรกิจต่างๆ ที่ผู้ใช้งานสนใจ ข้อมูลจะถูกรวบรวมมาจากระบบการดำเนินงานต่างๆ ที่หลากหลายรวมถึงแหล่งข้อมูลภายนอกด้วย ข้อมูลในคลังข้อมูลจะมีแกนเวลาเข้ามาเกี่ยวข้องเสมอ มีความละเอียดหลายระดับ และข้อมูลในคลังข้อมูลจะไม่ถูกเปลี่ยนแปลงจากผู้ใช้งานแต่อย่างใดก็ดี สำหรับคนที่ยังไม่มีประสบการณ์เกี่ยวกับคลังข้อมูลมากนักอาจได้ยินหรือรับรู้อีกสิ่งหนึ่ง นั่นคือ “**ดาต้ามาร์ท (data mart)**” และอาจเกิดความสับสนเกี่ยวกับความสัมพันธ์และความสอดคล้องระหว่าง “คลังข้อมูล” และ “ดาต้ามาร์ท” ก็เป็นไปได้หลายๆ คนจะมองว่าสองคำนี้เหมือนกันหรือสื่อถึงสิ่งเดียวกัน แต่แท้จริงแล้วทั้งสองคำนี้ไม่เหมือนกัน ซึ่งจากผู้ที่มีประสบการณ์กับคลังข้อมูลจะมองว่า



“คลังข้อมูล” คือ คลังข้อมูลของทั้งองค์กรหรือที่เรียกว่า **“enterprise data warehouse”** แต่สำหรับ **“ดาต้ามาร์ท”** คือ คลังข้อมูลส่วนย่อยๆ หรือคลังข้อมูลที่สนับสนุนการทำงานแต่ละแผนกขององค์กร

ซึ่งจากความแตกต่างในเรื่องของฟังก์ชันการทำงาน จำนวนข้อมูล ผู้ใช้งาน และปัจจัยอื่นๆ ระหว่างระบบทั้งสอง เราจะเป็นต้องพิจารณาสิ่งต่างๆ ดังต่อไปนี้

- ในการสร้างคลังข้อมูลเราควรใช้วิธีใดระหว่างการสร้างคลังข้อมูลแบบ top-down หรือ bottom-up?
กล่าวคือเราจะทำการสร้างคลังข้อมูลโดยพิจารณารายละเอียดทั้งหมดก่อน แล้วจึงทำการสร้างข้อมูลแต่ละส่วน หรือเราจะทำการสร้างคลังข้อมูลส่วนย่อยๆ ก่อน แล้วค่อยทำการรวมคลังข้อมูลเหล่านั้นให้เป็นคลังข้อมูลสำหรับทั้งองค์กร
- เราควรทำการสร้างคลังข้อมูลประเภทใดระหว่างคลังข้อมูลสำหรับทั้งองค์กรหรือคลังข้อมูลของแต่ละแผนก?
- เราควรสร้างอย่างใดอย่างหนึ่งก่อนระหว่างคลังข้อมูลสำหรับทั้งองค์กร หรือคลังข้อมูลของแต่ละแผนก?
- ถ้าเราทำการสร้างคลังข้อมูลของแต่ละแผนก เราควรสร้างคลังข้อมูลในลักษณะเป็นแบบที่เป็นดาต้ามาร์ทที่เชื่อมต่อกันหรือเป็นอิสระต่อกัน (Dependent or Independent data mart)?

จากปัจจัยข้างต้น จะมีคำถามอื่นๆ ตามมาอีกมากมาย เช่น เราจะต้องมองภาพรวมกว้างๆ ของทั้งองค์กรเพื่อทำการสร้างคลังข้อมูลแบบ top-down ใช่หรือไม่? หรือ เราควรจะเริ่มจากการสร้างคลังข้อมูลแบบ bottom-up โดยทำการพิจารณาความต้องการของแต่ละส่วนงาน/แผนกใช่หรือไม่? เราควรที่จะสร้างคลังข้อมูลขนาดใหญ่แล้วทำการจัดเก็บข้อมูลลงในแต่ละดาต้ามาร์ทหรือไม่? หรือเราควรที่จะสร้างแต่ละดาต้ามาร์ทแล้วค่อยทำการรวมดาต้ามาร์ทที่สร้างขึ้นให้เป็นคลังข้อมูลขนาดใหญ่? เราควรสร้างดาต้ามาร์ทให้เป็นอิสระต่อกันหรือไม่? หรือเราควรสร้างดาต้ามาร์ทให้มีความเกี่ยวเนื่องกันไหม? คำถามเหล่านี้เป็นคำถามที่สำคัญและส่งผลกระทบต่อประมาณการสร้างคลังข้อมูล ส่งผลกระทบต่อระยะเวลาในการสร้างคลังข้อมูล และส่งผลกระทบต่อรูปแบบการใช้งานคลังข้อมูล



ดังนั้นก่อนที่จะทำการสร้างคลังข้อมูลเราควรจะต้องพิจารณาให้รอบคอบว่าเราควรที่จะสร้างคลังข้อมูลในลักษณะใด โดยใช้วิธีการอะไร เพื่อควบคุมค่าใช้จ่าย เวลาในการดำเนินงาน และฟังก์ชันการทำงานที่ครบครันที่สุด แต่ก่อนที่จะตอบปัญหาต่างๆ ข้างต้น ลองพิจารณารูปที่ 2-7 เพื่อทราบถึงความแตกต่างระหว่าง **“data warehouse”** และ **“data mart”** เพื่อที่จะได้ทำการออกแบบหรือกำหนดการสร้างคลังข้อมูลต่อไป

Monthly Summary

- Corporate/Enterprise-wide
- Union of all data marts
- Data received from staging area
- Queries on presentation resource
- Structure for corporate view of data
- Organized on E-R model

Quarterly Summary

- Departmental
- A single business process
- STARjoin (facts & dimensions)
- Technology optimal for data access and analysis
- Structure to suit the departmental view of data

รูปที่ 2-7 ความแตกต่างระหว่างคลังข้อมูลและดาต้ามาร์ท

วิธีการสร้างคลังข้อมูล



หลังจากที่เราทราบถึงความแตกต่างของ “Data warehouse” และ “Data mart” แล้ว เราควรที่จะพิจารณาถึงวิธีการสร้างทั้งสองระบบ ที่จะสามารถจำแนกวิธีได้สร้างได้ 2 วิธี ดังนี้



1

Top-down

เป็นวิธีการสร้างคลังข้อมูลที่ถูกเสนอโดย “Bill Inmon” ซึ่งได้นิยามคลังข้อมูลที่ถูกสร้างโดยวิธีการนี้ว่าเป็น “ศูนย์กลางคลังข้อมูลสำหรับองค์กร” ที่มีการจัดเก็บข้อมูลที่มีความละเอียดค่อนข้างสูงและมีการทำนอร์มอลไลซ์กับข้อมูล โดยคลังข้อมูลที่สร้างขึ้นจากวิธีการ top-down จะอยู่ที่ศูนย์กลางที่มีการสร้าง “logical framework” สำหรับสนับสนุนการทำธุรกิจอย่างชาญฉลาดขององค์กร

ข้อดี

ข้อดีของวิธีการ *Top-down* จะประกอบไปด้วย

- สามารถมองข้อมูลได้ทั่วทั้งองค์กร
- สถาปัตยกรรมเป็นเนื้อเดียวกันและไม่ได้เป็นแบบการรวมกันของหลายๆ ดาต้ามาร์ท
- ทำการเก็บข้อมูลไว้ที่เดียว
- มีการควบคุมและกำหนดกฎเกณฑ์ต่างๆ จากศูนย์กลาง

ข้อเสีย

ข้อเสียของวิธีการ *Top-down* จะประกอบไปด้วย

- ใช้เวลาในการสร้างค่อนข้างนาน
- มีความเสี่ยงที่เกิดความล้มเหลวค่อนข้างสูง
- ต้องการผู้สร้างที่มีความรู้ ความสามารถสูงในการที่จะสร้างการเชื่อมโยงฟังก์ชันการทำงานที่มีการข้ามสายงาน
- เสียค่าใช้จ่ายค่อนข้างมาก

จากข้างต้นเราจะสามารถเห็นภาพกว้างๆ ของวิธีการสร้างคลังข้อมูลแบบ Top-down ที่จะทำให้เราได้ข้อมูลเป็นกลุ่มก้อนเป็นชั้นเดียวกัน แต่อย่างไรก็ดีการสร้างวิธีนี้จะใช้เวลานาน เนื่องจากต้องทำความเข้าใจเกี่ยวกับการดำเนินธุรกิจทั้งองค์กร มีความเสี่ยงสูงที่จะเกิดความล้มเหลวหากทีมผู้สร้างคลังข้อมูลยังไม่มีประสบการณ์เกี่ยวกับการสร้างคลังข้อมูลเพียงพอ และมีความต้องการผู้สร้างที่มีความเชี่ยวชาญทั้งในเชิงเทคนิคและเชิงธุรกิจสูง

2

Bottom-up

เป็นวิธีการสร้างคลังข้อมูลที่ถูกเสนอโดย Ralph Kimball ซึ่งได้นิยามคลังข้อมูลที่ถูกสร้างโดยวิธีการนี้ว่าเป็น “กลุ่มของดาต้ามาร์ทที่สอดคล้องกัน” โดยปัจจัยหลักของวิธีการนี้จะอยู่ที่ความสอดคล้องกันของดาต้ามาร์ทที่สร้างขึ้นเพื่อสนับสนุนการวิเคราะห์ข้อมูลในแต่ละส่วนงาน การสร้างคลังข้อมูลด้วยวิธีการ bottom-up จะเริ่มจากการสร้างดาต้ามาร์ทของแต่ละส่วนงานเพื่อให้ผู้ใช้หรือพนักงานในแต่ละแผนกสามารถทำการวิเคราะห์ข้อมูลและสร้างรายงานในแง่มุมต่างๆ ที่สอดคล้องกับการดำเนินธุรกิจของแผนกนั้นๆ ได้ ดาต้ามาร์ทแต่ละส่วนจะมีการเก็บข้อมูลที่มีความละเอียดสูง รวมถึงข้อมูลที่เป็นผลสรุปตามความต้องการในการวิเคราะห์ข้อมูลของผู้ใช้ เมื่อทำการสร้างดาต้ามาร์ทหลายๆ ดาต้ามาร์ทแล้ว จากนั้นจะทำการเชื่อมโยงทุกๆ ดาต้ามาร์ทเข้าด้วยกันโดยคำนึงถึงความสอดคล้องของข้อมูลเป็นหลัก

ข้อดี

ข้อดีของวิธีการ *Bottom-up* จะประกอบไปด้วย

- สามารถการดำเนินการได้เร็วขึ้นและง่ายขึ้น โดยการพิจารณาข้อมูลแต่ละส่วนงาน
- มีความเสี่ยงของความล้มเหลวน้อย
- สามารถกำหนดให้ส่วนงาน/แผนกที่มีความสำคัญค่อนข้างมากสามารถการสร้างดาต้ามาร์ทได้ก่อน
- ช่วยให้ผู้สร้างสามารถทำการเรียนรู้ที่ละส่วนงาน

ข้อเสีย

ข้อเสียของวิธีการ *Bottom-up* จะประกอบไปด้วย

- แต่ละดาต้ามาร์ทจะมีข้อมูลของตัวเองเท่านั้นซึ่งเป็นข้อมูลที่ค่อนข้างแคบ
- อาจทำให้เกิดความซ้ำซ้อนของข้อมูล อาจต้องทำการเก็บข้อมูลหนึ่งๆ ไว้ในทุกดาต้ามาร์ท
- การสร้างแต่ละดาต้ามาร์ทก่อน แล้วค่อยรวมกันอาจทำให้ข้อมูลไม่สอดคล้องกัน
- อาจทำให้อินเทอร์เฟซ (interface) ของทุกๆ ดาต้ามาร์ทนั้นไม่สอดคล้องกัน และเมื่อทำการรวมดาต้ามาร์ทเข้าด้วยกัน อาจไม่สามารถจัดการให้อินเทอร์เฟซต่างๆ ให้เป็นมาตรฐานเดียวกันได้

วิธีการสร้างคลังข้อมูลแบบ Bottom-up จะทำการสร้างดาต้ามาร์ททีละส่วน โดยที่เราสามารถกำหนดความสำคัญของส่วนงานเพื่อกำหนดว่าดาต้ามาร์ทใดควรจะมีการสร้างก่อน แต่ข้อเสียที่ชัดเจนที่สุดคือ ข้อมูลจะกระจายออกเป็นส่วนๆ ไม่รวมเป็นกลุ่มก้อน ซึ่งเมื่อดาต้ามาร์ทเป็นอิสระต่อกันจะทำให้ไม่สามารถมองเห็นความต้องการทั้งหมดทั่วทั้งองค์กรได้

SECTION 7

แนวปฏิบัติสำหรับการสร้าง คลังข้อมูล

จากวิธีการสร้างคลังข้อมูลทั้งสองวิธีข้างต้น เราจะทราบว่าแต่ละวิธีจะมีข้อดีและข้อเสียที่แตกต่างกัน ดังนั้นในการสร้างคลังข้อมูลเราจะต้องพิจารณาว่าสิ่งที่เราต้องการคืออะไร? องค์กรของเราต้องการที่จะมองหาผลลัพธ์ระยะยาวหรือดาต้ามาร์ทที่มีข้อมูลไม่มากที่สามารถสร้างได้รวดเร็ว ณ ปัจจุบัน? องค์กรของเราต้องการใช้เวลาในการสร้างระยะสั้นหรือไม่? หรือองค์กรของเรากำลังมองหาวิธีการที่สามารถใช้งานได้จริง ซึ่งจากวิธีการสร้างแบบ Top-down และ Bottom-up ต่างก็มีข้อเสีย ดังนั้นวิธีการที่น่าจะดีที่สุดก็คือ“การรวมกันของทั้งสองวิธี” ที่จะทำให้เราสามารถเห็นภาพกว้างๆ ของทั้งองค์กร โดยในการวางแผนการสร้างสำหรับทั้งองค์กรจะใช้วิธีการแบบ top-down แต่เราจะประยุกต์ใช้วิธีการแบบ bottom-up ในการสร้างแต่ละดาต้ามาร์ทที่มีความเหมาะสม โดยทำการกำหนดลำดับความสำคัญของแต่ละส่วนงานที่จะทำการสร้างคลังข้อมูล

ซึ่งขั้นตอนวิธีในการสร้างคลังข้อมูลที่น่าวิธีการ top-down และ bottom-up มารวมกันจะประกอบไปด้วย 4 ขั้นตอน ดังนี้

- 1 วางแผนและกำหนดความต้องการของทั้งองค์กรทุกๆ ระดับ ตั้งแต่ความต้องการของนักวิเคราะห์ข้อมูล ผู้จัดการ ผู้บริหาร กรรมการผู้จัดการ และ อื่นๆ
- 2 สร้างสถาปัตยกรรมโดยรวมสำหรับคลังข้อมูลที่สมบูรณ์
- 3 กำหนดวิธีการที่จะทำให้ข้อมูลที่ถูกเก็บอยู่ในคลังข้อมูลและแต่ละดาต้ามาร์ทมีความสอดคล้องกัน และเป็นมาตรฐานเดียวกัน
- 4 ทำการสร้างคลังข้อมูลทีละส่วนงานเรียงต่อกัน โดยทำการสร้างดาต้ามาร์ทหนึ่งๆ ณ ช่วงเวลาหนึ่งๆ เท่านั้น

จากแนวปฏิบัติทั้ง 4 ข้อจะทำให้เราสามารถมองได้ว่าคลังข้อมูลคือ “กลุ่มของดาต้ามาร์ทที่สอดคล้องกัน” โดยที่แต่ละดาต้ามาร์ทจะให้บริการการดำเนินธุรกิจของแต่ละส่วนงานและภาพรวมของทั้งองค์กรด้วย ซึ่งเราสามารถเรียกกลุ่มของดาต้ามาร์ทที่ทำการสร้างขึ้นว่าเป็น “Enterprise data warehouse” โดยแนวปฏิบัติในการสร้างคลังข้อมูลนั้นจะมีหัวใจหลักอยู่ที่การวางแผนและกำหนดความต้องการของทั้งองค์กร ซึ่งเราจะต้องเก็บรวบรวมความต้องการให้ได้ทั้งหมดจากผู้ใช้ทุกระดับ และการทำให้ข้อมูลมีความสอดคล้องกัน ทั้งในส่วนของ ชนิดของข้อมูล ความยาวของฟิลด์ ความถูกต้องแม่นยำ และความหมายที่สื่อไปในทางเดียวกัน เพื่อหลีกเลี่ยงความแตกต่างของข้อมูล หรือความไม่สอดคล้องของข้อมูลระหว่างดาต้ามาร์ทที่อาจจะทำให้เกิดความไม่ถูกต้องของข้อมูลเกิดขึ้น

คำถามท้ายบท



1. นิยามหรือคำจำกัดความของ “คลังข้อมูล” คืออะไร
2. จงอธิบายและแจกแจงคุณลักษณะเด่นของคลังข้อมูล
3. เพราะเหตุใดทุกๆ คลังข้อมูลจะต้องมีแกนเวลาเข้ามาเกี่ยวข้อง
4. จงอธิบายหรือให้คำจำกัดความของ “ข้อมูลที่มีความละเอียดหลายระดับ”
5. คลังข้อมูลประกอบไปด้วยส่วนประกอบอะไรบ้าง
6. คลังข้อมูลและดาต้ามาร์ท เหมือนหรือแตกต่างกันอย่างไร
7. เพราะเหตุใดคลังข้อมูลจึงจำเป็นต้องมีหลายแหล่งข้อมูล และแหล่งข้อมูลของคลังข้อมูลมีกี่ชนิดอะไรบ้าง
8. เพราะเหตุใดเราจึงต้องเรียกใช้ “data staging” หรือ “staging area”
9. การสร้างคลังข้อมูลแบบ top-down และ bottom-up มีความแตกต่างกันอย่างไร
10. ในการสร้างคลังข้อมูล เราควรเลือกใช้วิธีการสร้างแบบใด เพราะเหตุใด