การพัฒนาระบบคลังข้อมูล

Building a Data Warehouse

เบญจมาศ เต็มอุดม [1] และ ดร.ภัทรชัย ลลิตโรจน์วงศ์ [2]

บทคัดย่อ

ดาต้าแวร์เฮ้าส์ หรือ คลังข้อมูล คือฐานข้อมูลขนาดใหญ่ที่รวบรวมข้อมูลทั้งจากแหล่งข้อมูลภายในและภายนอกองค์กร โดยมีรูปแบบและวัตถุประสงค์ของการจัดเก็บข้อมูลแตกต่างจากฐานข้อมูลปฏิบัติการทั่วไป โดยข้อมูลในคลังข้อมูลมีการจัดเก็บเฉพาะเนื้อหาในเรื่องที่สนใจ (subject oriented) ในลักษณะการรวมเป็นหนึ่งเดียว (integration) และข้อมูลมีความสัมพันธ์กับระยะเวลา (time variancy) นอกจากนี้ ข้อมูลจะไม่มีการเปลี่ยนแปลงง่าย ๆ (nonvolatile) ทั้งนี้ เพื่อวัตถุประสงค์ในการนำไปเป็นข้อมูลเพื่อการตัดสินใจบริหารงานองค์กรของผู้บริหาร อย่างไรก็ตาม สิ่งที่สำคัญมากในการสร้างคลังข้อมูลคือเครื่องมือที่จะนำเข้ามาจัดการกับข้อมูลจากฐานข้อมูลหลักภายในและภายนอก เรียกว่า data warehouse tool ซึ่งทำหน้าที่ในการดึงข้อมูลที่ต้องการ เปลี่ยนรูปแบบให้เป็นแบบเดียวกัน และนำเข้าสู่คลังข้อมูลเพื่อให้เป็นข้อมูลที่เหมาะสมที่จะนำไปวิเคราะห์ต่อไป

Abstract

A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile, collection of data. It collects data from sources both inside and outside organizations to support management decision processes. Data warehouses are different from operational (or production) systems in terms of formats, purposes of use, and processes. The key issue that one wants to set up data warehouse to be concerned is data warehouse tools. The data warehouse tools can be used to extract meaning from informational assets, format and load this meaning information into the data warehouse. If these process are performed properly, the warehouse data will be ready for analysis.

1.บทนำ

เนื่องจากสภาพเศรษฐกิจปัจจุบันที่ยังไม่จัดว่าพ้นภาวะวิกฤต ธุรกิจหลายประเภทจึงยังต้องการการวิเคราะห์วางแผน และตัดสินใจอย่างถูกต้อง รวดเร็ว เพื่อช่วยให้ธุรกิจสามารถดำเนินไปได้ ดังนั้นข้อมูลจึงเป็นปัจจัยที่สำคัญยิ่งยวดต่อการดำเนินการนั้น การใช้ข้อมูลเป็นเครื่องมือสำคัญในการตัดสินใจลงทุนทางธุรกิจ และวางแผนกลยุทธ์ทางการตลาดเพื่อแข่งขันกับคู่แข่งทางการค้า ฉะนั้นก็อาจกล่าวได้ว่า การมีข้อมูลมากทำให้มีโอกาสและมีชัยเหนือคู่แข่งในระดับหนึ่ง

แต่ทว่าหากมองในทางกลับกัน การมีข้อมูลจำนวนมากแต่ขาดการจัดเรียงให้เป็นระบบระเบียบ ยุ่งยากในการเข้าถึงและค้นคืน ธุรกิจอาจต้องเสียค่าใช้จ่ายจำนวนมหาศาลในการเก็บรักษาข้อมูลเหล่านั้นไว้โดยไม่จำเป็น เพราะไม่ได้รับประโยชน์จากข้อมูลที่มี นอกจากนี้หากมีการนำข้อมูลมาวิเคราะห์อย่างผิดพลาดอาจจะก่อให้เกิดผลเสียหายได้ ซึ่งเป็นการสูญเสียโอกาสทางธุรกิจไป เพราะฉะนั้นในยุคที่ผู้บริหารมีความต้องการใช้ข้อมูลเพื่อการตัดสินใจมากขึ้น การจัดระบบระเบียบข้อมูล เพื่อนำเสนอข้อมูลที่มีคุณค่าและผ่าน

---

[1] นักศึกษาปริญญาโทคณะเทคโนโลยีสารสนเทศ สจล.

[2] อาจารย์คณะเทคโนโลยีสารสนเทศ สจล.

การกลั่นกรองแล้วแก่ผู้บริหาร เพื่อใช้ในการตัดสินใจให้ทันต่อเหตุการณ์จึงเป็นสิ่งที่จำเป็นอย่างยิ่ง

แนวความคิดของการสร้างคลังข้อมูลจึงเกิดขึ้นเพื่อเป็นที่เก็บรวบรวมข้อมูลสำคัญและจำเป็นจากแหล่งต่าง ๆ ซึ่งเป็นประโยชน์ต่อการตัดสินใจของผู้บริหาร เพื่อให้ผู้บริหารสามารถเรียกใช้ข้อมูลที่ต้องการได้อย่างรวดเร็วและมีประสิทธิภาพมากขึ้น ข้อมูลเชิงบริหารนี้จะสามารถช่วยลดปัญหาที่เกิดจากการใช้ข้อมูลจากฐานข้อมูลปฏิบัติการ (operational database) ซึ่งเป็นการเก็บข้อมูลในรูปแบบ transaction system ได้ ซึ่งโดยทั่วไปปัญหาที่พบเมื่อต้องการข้อมูลที่ช่วยในการตัดสินใจ ได้แก่ [1]

- การเรียกข้อมูลจากฐานข้อมูลปฏิบัติการ ซึ่งมีขนาดใหญ่ ทำให้ประสิทธิภาพของระบบลดลงและทำงานได้ช้าลง
- ข้อมูลที่นำเสนอมีรูปแบบเดียว ไม่สามารถเปลี่ยนแปลงได้ตามความต้องการของผู้บริหาร
- ไม่สามารถหาคำตอบในเชิงพยากรณ์ได้
- ไม่ตอบสนองการทำคิวรีที่ซับซ้อนได้ดีเท่าที่ควร
- ข้อมูลถูกจัดเก็บอยู่ตามฐานข้อมูลของระบบงานต่าง ๆ ซึ่งยากแก่การเรียกใช้และขาดความสัมพันธ์ทางธุรกิจ

2.แนวคิดเกี่ยวกับคลังข้อมูล

2.1 นิยามของคลังข้อมูล

คลังข้อมูล หมายถึง ฐานข้อมูลขนาดใหญ่ขององค์กรหรือหน่วยงานหนึ่ง ๆ ซึ่งเก็บรวบรวมข้อมูลจากฐานข้อมูลระบบงานประจำวัน หรือเรียกอีกอย่างว่า operational database และฐานข้อมูลอื่นภายนอกองค์กร หรือเรียกว่า external database โดยข้อมูลที่ถูกจัดเก็บในคลังข้อมูลนั้น มีวัตถุประสงค์ในการนำมาใช้งานและมีลักษณะของการจัดเก็บแตกต่างไปจากข้อมูลในฐานข้อมูลระบบงานอื่น โดยข้อมูลในคลังข้อมูลจะถูกนำมาใช้เพื่อสนับสนุนการตัดสินใจบริหารงานของผู้บริหาร โดยเฉพาะการเป็นข้อมูลพื้นฐานให้กับระบบงานเพื่อการบริหารงานอื่น เช่น ระบบ DSS และระบบ CRM เป็นต้น [2, 5, 6]

2.2 คุณลักษณะเฉพาะของคลังข้อมูล

จากนิยามของคลังข้อมูลที่บอกถึงความแตกต่างกันระหว่างคลังข้อมูลกับฐานข้อมูลปฏิบัติการ ซึ่งสามารถสรุปคุณลักษณะของคลังข้อมูล [2] ได้ดังนี้

1. Subject oriented หรือการแบ่งโครงสร้างตามเนื้อหา หมายถึง คลังข้อมูลถูกออกแบบมาเพื่อมุ่งเน้นไปในแต่ละเนื้อหาที่สนใจ ไม่ได้เน้นไปที่การทำงานหรือกระบวนการแต่ละอย่างโดยเฉพาะเหมือนอย่างฐานข้อมูลปฏิบัติการ ในส่วนของรายละเอียดข้อมูลที่จะจัดเก็บในระบบทั้งสองแบบก็จะแตกต่างกับไปตามความต้องการใช้งานด้วยเช่นกัน คลังข้อมูลจะไม่จัดเก็บข้อมูลที่ไม่มีส่วนเกี่ยวข้องกับการประมวลผลเพื่อสนับสนุนการตัดสินใจ ในขณะที่ข้อมูลนั้นจะถูกเก็บไว้ในฐานข้อมูลปฏิบัติการหากมีส่วนเกี่ยวข้องกับกระบวนการทำงาน
2. Integration หรือการรวมเป็นหนึ่งเดียว ซึ่งถือได้ว่าเป็นคุณลักษณะที่สำคัญที่สุดของคลังข้อมูล คือ การรวบรวมข้อมูลจากหลายฐานข้อมูลปฏิบัติการเข้าด้วยกัน และทำให้ข้อมูลมีมาตรฐานเดียวกัน เช่นกำหนดให้มีค่าตัวแปรของข้อมูลในเนื้อหาเดียวกันให้เป็นแบบเดียวกันทั้งหมด
3. Time variancy หรือความสัมพันธ์กับเวลา หมายถึง ข้อมูลในคลังข้อมูลจะต้องจัดเก็บโดยกำหนดช่วงเวลาเอาไว้ โดยจะสัมพันธ์กับการดำเนินธุรกิจของหน่วยธุรกิจนั้น เพราะในการตัดสินใจในการบริหารจำเป็นต้องมีข้อมูลเปรียบเทียบในแต่ละช่วงเวลา แต่ละจุดของข้อมูลจะเกี่ยวข้องกับจุดของเวลาและข้อมูลแต่ละจุดสามารถเปรียบเทียบกันได้ตามแกนของเวลา
4. Nonvolatile หรือความเสถียรของข้อมูล หมายถึง ข้อมูลในคลังข้อมูลจะไม่ถูกเปลี่ยนแปลงง่าย ๆ ไม่ว่าจะเป็นการเพิ่มเติมข้อมูลใหม่ หรือการปรับปรุงแก้ไขข้อมูลเดิมที่บรรจุอยู่แล้ว ผู้ใช้ทำได้เพียงการเข้าถึงข้อมูลเท่านั้น

3. สถาปัตยกรรมคลังข้อมูล (Data Warehouse Architecture - DWA)

DWA เป็นโครงสร้างมาตรฐานที่ใช้อธิบายเพื่อให้เข้าใจแนวคิด และกระบวนการของคลังข้อมูลนั้น ๆ ซึ่งโดยทั่วไปแล้ว คลังข้อมูลแต่ละระบบอาจจะมีรูปแบบที่ไม่เหมือนกันได้เพื่อให้เหมาะสมกับองค์กรนั้น ๆ ทั้งนี้ส่วนประกอบต่าง ๆ ภายใน DWA ที่สำคัญ ได้แก่ [4]

1. Operational database หรือ external database layer ทำหน้าที่จัดการกับข้อมูลในระบบงานปฏิบัติการ หรือแหล่งข้อมูลภายนอกองค์กร

2. Information access layer เป็นส่วนที่ผู้ใช้ปลายทางติดต่อผ่านโดยตรง ประกอบด้วยฮาร์ดแวร์และซอฟต์แวร์ที่ใช้ในการแสดงผลเพื่อการวิเคราะห์ โดยมีเครื่องมือช่วยเป็นตัวกลางที่ผู้ใช้ใช้ติดต่อกับคลังข้อมูล โดยในปัจจุบันเครื่องมือที่ได้รับความนิยมเพิ่มขึ้นอย่างรวดเร็ว นั่นคือ Online Analytical Processing Tool หรือ OLAP tool ซึ่งเป็นเครื่องมือที่มีความสามารถในการวิเคราะห์ที่ซับซ้อนและแสดงข้อมูลในรูปแบบหลายมิติ

3. Data access layer เป็นส่วนต่อประสานระหว่าง Information access layer กับ operational layer

4. Data director (metadata) layer เพื่อให้เข้าถึงข้อมูลได้ง่ายขึ้น และเป็นการเพิ่มความเร็วในการเรียกและดึงข้อมูลของคลังข้อมูล

5. Process management layer ทำหน้าที่จัดการกระบวนการทำงานทั้งหมด

6. Application messaging layer เป็นมิดเดิลแวร์ทำหน้าที่ในการส่งข้อมูลภายในองค์กรผ่านทางเครือข่าย

7. Data warehouse (physical) layer เป็นแหล่งเก็บข้อมูลของทั้ง information data และ external data ในรูปแบบที่ง่ายแก่การเข้าถึงและยืดหยุ่นได้

8. Data staging layer เป็นกระบวนการแก้ไข และดึงข้อมูลจาก external database

4. เทคนิคในการสร้างคลังข้อมูล

4.1 การเคลื่อนที่ของข้อมูลในคลังข้อมูล

ข้อมูลที่จะจัดเก็บภายในคลังข้อมูลมีการเคลื่อนที่ของข้อมูล (information flow) 5 ประเภท [8] ดังนี้

1. Inflow คือการนำข้อมูลจากฐานข้อมูลอื่นเข้าสู่คลังข้อมูล ทั้งฐานข้อมูลภายในและภายนอกองค์กร โดยในขั้นนี้อาจมีการเปลี่ยนแปลงโครงสร้างข้อมูล การทำ denormalize การลบหรือเพิ่มฟิลด์เพื่อให้ข้อมูลทั้งหมดอยู่ในเนื้อหาที่สนใจเดียวกัน ในขั้นตอนนี้อาจใช้เครื่องมือที่เรียกว่า data warehouse tool

2. Upflow เมื่อข้อมูลที่เราต้องการอยู่ในคลังข้อมูลแล้ว ในบางครั้งอาจต้องมีการเพิ่มคุณค่าให้กับข้อมูลด้วยเพื่อให้ข้อมูลอยู่ในรูปแบบที่เป็นประโยชน์มากที่สุดต่อการนำเครื่องมือมาใช้ ซึ่งได้แก่การจัดกลุ่มข้อมูล หาค่าทางสถิติที่ซับซ้อน จัดข้อมูลให้อยู่ในรูปแบบหรือเทมเพลตมาตรฐาน

3. Downflow เป็นขั้นตอนของการปรับปรุงเปลี่ยนแปลงข้อมูลที่เก่าและไม่อยู่ในเนื้อหาที่องค์กรสนใจอีกต่อไปออกไปจากคลังข้อมูลขององค์กร

4. Outflow เป็นขั้นที่ผู้ใช้เรียกใช้ข้อมูลในคลังข้อมูลผ่านเครื่องมือต่าง ๆ โดยการเรียกใช้อาจมีเพียงขอดูเป็นครั้งคราว เป็นประจำทุกวัน/เดือน หรือแม้กระทั่งต้องการแบบทันที

5. Metafolw ข้อมูลที่จัดเก็บในคลังข้อมูลจะถูกทำข้อมูลไว้อีกชุดหนึ่ง เป็นแหล่งที่มาของข้อมูลนั้น หรือแม้กระทั่งที่อยู่ของข้อมูลนั้นในคลังข้อมูลและข้อมูลอื่นที่เกี่ยวข้อง

4.2 วิธีการออกแบบฐานข้อมูลสำหรับคลังข้อมูล

วิธีการนี้ถูกเสนอโดย Kimball ในปี 1996 เรียกว่า “Nine-Step Methodology” [7] โดยวิธีการนี้เริ่มจากการออกแบบจากส่วนย่อยที่แสดงถึงแต่ละระบบงานขององค์กร หรือเรียกอีกอย่างหนึ่งว่าดาต้ามาร์ท (data mart) โดยเมื่อออกแบบแต่ละส่วนสำเร็จแล้ว จึงนำมารวมกันเป็นคลังข้อมูล

ขององค์กรในขั้นสุดท้าย ซึ่งขั้นตอนทั้ง 9 ขั้นตอน มีรายละเอียดดังนี้

1. กำหนดดาต้ามาร์ท คือการเลือกว่าจะสร้างดาต้ามาร์ทของระบบงานใดบ้าง และระบบงานใดเป็นระบบงานแรก โดยองค์กรจะต้องสร้าง E-R model ที่รวมระบบงานทุกระบบขององค์กรไว้ แสดงความเชื่อมโยงของแต่ละระบบงานอย่างชัดเจน และสิ่งที่ต้องคำนึงถึงในการเลือกระบบงานที่จะเป็นดาต้ามาร์ทแรกนั้น มี 3 ปัจจัยที่เกี่ยวข้อง ได้แก่ จะต้องสามารถพัฒนาออกมาได้ทันตาม*เวลา*ที่ต้องการ โดยอยู่ใน*งบประมาณ*ที่กำหนดไว้ และต้องตอบ*ปัญหาทางธุรกิจ*ให้แก่องค์กรได้ ดังนั้น ดาต้ามาร์ทแรกควรจะเป็นของระบบงานที่นำรายได้เข้ามาสู่องค์กรได้ เช่น ระบบงานขาย เป็นต้น

2. กำหนด fact table ของดาต้ามาร์ท คือการกำหนดเนื้อหาหลักที่ควรจะเป็นของดาต้ามาร์ท โดยการเลือกเอนทิตีหลักและกระบวนการที่เกี่ยวกับเอนทิตีนั้น ๆ ออกมาจาก E-R model ขององค์กร นั่นหมายถึงว่าจะทำให้เราทราบถึง dimension table ที่ควรจะมีด้วย

3. กำหนดแอตทริบิวต์ที่จำเป็นในแต่ละ dimension table คือการกำหนดแอตทริบิวต์ที่บอกหรืออธิบายรายละเอียดของ dimension ได้ ทั้งนี้ แอตทริบิวต์ที่เป็น primary key ควรเป็นค่าที่คำนวณได้ กรณีที่มีดาต้ามาร์ทมากกว่าหนึ่งดาต้ามาร์ทมี dimension เหมือนกัน นั่นหมายถึงว่าแอตทริบิวต์ใน dimension นั้นจะต้องเหมือนกันทุกประการ แต่นั่นก็ไม่อาจจะแก้ไขปัญหาการจัดเก็บข้อมูลซ้ำซ้อน อันจะนำมาสู่ความแตกต่างกันของข้อมูลชุดเดียวกัน ปัญหานี้จึงเป็นการดีที่จะมีการใช้ dimension table ร่วมกันในแต่ละ fact table ที่จำเป็นต้องมี dimension ดังกล่าว โดยเรียก dimension table ลักษณะแบบนี้ว่า conformed และเรียก fact table ว่า fact constellation เราสามารถกำหนดข้อดีของการใช้ dimension table ร่วมกันได้ดังนี้

(1) แน่ใจได้ว่าในแต่ละรายงานจะออกมาสอดคล้องกัน

(2) สามารถสร้างดาต้ามาร์ทในเวลาต่าง ๆ กันได้

(3) สามารถเข้าถึงดาต้ามาร์ทโดยผู้พัฒนากลุ่มอื่น ๆ

(4) สามารถรวบรวมดาต้ามาร์ทหลาย ๆ อันเข้าด้วยกัน

(5) สามารถออกแบบคลังข้อมูลร่วมกันได้

4. กำหนดแอตทริบิวต์ที่จำเป็นใน fact table โดยแอตทริบิวต์หลักใน fact table จะมาจาก primary key ในแต่ละ dimension table นอกจากนี้แล้ว ยังสามารถมีแอตทริบิวต์ที่จำเป็นอื่น ๆ ประกอบอยู่ด้วย เช่น แอตทริบิวต์ที่ได้จากการคำนวณค่าเบื้องต้นที่จำเป็นสำหรับการคงอยู่ของแอตทริบิวต์อื่นใน fact table เรียกอีกอย่างหนึ่งว่า measure การกำหนดแอตทริบิวต์นี้ไม่ควรจะเลือกแอตทริบิวต์ที่คำนวณค่าไม่ได้ เช่น เป็นตัวหนังสือหรือไม่ใช่ตัวเลข เป็นต้น และไม่ควรเลือกแอตทริบิวต์ที่ไม่เกี่ยวข้องกับเนื้อหาของ fact table ที่เราสนใจด้วย

5. จัดเก็บค่าการคำนวณเบื้องต้นใน fact table คือการจัดเก็บค่าที่ได้จากการคำนวณให้เป็นแอตทริบิวต์หนึ่งใน fact table ถึงแม้ว่าจะสามารถหาค่าได้จากแอตทริบิวต์อื่น ๆ ก็ตาม ทั้งนี้เพื่อให้การสอบถามมีประสิทธิภาพมากขึ้น สามารถทำงานด้วยความเร็วที่เพิ่มขึ้น เนื่องจากไม่ต้องคำนวณค่าใหม่ทั้งหมด ถึงแม้ว่าจะเกิดความซ้ำซ้อนของข้อมูลในการจัดเก็บบ้างก็ตาม

6. เขียนคำอธิบายของ dimension table ทั้งนี้ก็เพื่อให้ผู้ใช้สามารถใช้งานดาต้ามาร์ทได้อย่างมีประสิทธิภาพ เพราะเกิดความเข้าใจอย่างดีในส่วนต่าง ๆ

7. กำหนดระยะเวลาในการจัดเก็บข้อมูลในฐานข้อมูล โดยอาจจะเป็นการจัดเก็บเพียงช่วงระยะเวลา 1-2 ปี หรือนานกว่านั้น ขึ้นอยู่กับความต้องการขององค์กร เนื่องจากองค์กรแต่ละประเภทมีความต้องการในการจัดเก็บข้อมูลต่างช่วงเวลากัน ทั้งนี้ขึ้นอยู่กับความจำเป็นหรือข้อกำหนดในการดำเนินธุรกิจ มีข้อสังเกตอยู่ 2 ประการที่น่าสนใจและสำคัญสำหรับการออกแบบแอตทริบิวต์ในเรื่องของการจัดเก็บข้อมูล ดังนี้

(1) ข้อมูลที่ถูกจัดเก็บไว้นานเกินไปมักเกิดปัญหาการอ่านหรือแปลข้อมูลนั้น ๆ จากแฟ้มหรือเทปเก่า

(2) เมื่อมีการนำรูปแบบเก่าของ dimension table มาใช้ อาจเกิดปัญหาการเปลี่ยนแปลงของ dimension อย่างช้า ๆ ได้

8. การติดตามปัญหาการเปลี่ยนแปลงของ dimension อย่างช้า ๆ คือ การเปลี่ยนเอาแอตทริบิวต์ของ dimension table เก่ามาใช้แล้วส่งผลกระทบต่อข้อมูลปัจจุบันของ dimension table โดยสามารถแบ่งประเภทของปัญหาที่เกิดได้เป็น 3 ประเภท ดังนี้

(1) เกิดการเขียนทับข้อมูลใหม่โดยข้อมูลเก่า

(2) เกิดเรคอร์ดใหม่ ๆ ขึ้นใน dimension

(3) เกิดเรคอร์ดที่มีทั้งค่าเก่าและใหม่ปนกันไป

9. กำหนดคิวรีเป็นการออกแบบด้านกายภาพเพื่อให้ผู้ใช้เกิดความสะดวกในการใช้งานและสามารถทำงานได้อย่างมีประสิทธิภาพ

เมื่อดำเนินการทั้ง 9 ขั้นตอนสำหรับแต่ละดาต้ามาร์ทเสร็จแล้ว จึงจะนำทั้งหมดมารวมกันเป็นภาพของคลังข้อมูลขององค์กรต่อไป

4.3 การแปลงข้อมูลเข้าสู่ดาต้ามาร์ท

เมื่อเราออกแบบฐานข้อมูลสำหรับแต่ละดาต้ามาร์ทเสร็จแล้ว ขั้นตอนต่อไปที่สำคัญยิ่งก็คือการนำข้อมูลจากแหล่งข้อมูลไปแปลงให้อยู่ในแพลตฟอร์มของฐานข้อมูลที่ได้ออกแบบไว้ นั่นก็คือการแปลงข้อมูล หรือ Extraction Transformation and Loading (ETL) นั่นเอง โดยที่คุณภาพของการแปลงข้อมูลเป็นสิ่งที่สำคัญมากสำหรับการสร้างคลังข้อมูล ความซับซ้อนของการแปลงข้อมูลและโครงสร้างของข้อมูลจะแตกต่างกันไปตามคลังข้อมูลที่แต่ละองค์กรต้องการ โดยที่การแปลงข้อมูลหมายรวมตั้งแต่การวิเคราะห์แหล่งข้อมูล กำหนดการ map ข้อมูล รวบรวมหรือสร้างข้อมูลภายนอก วางแผนและสร้างรูทีนของการแปลงข้อมูลและตรวจสอบความถูกต้องของข้อมูลที่ได้ สามารถสรุปเป็นขั้นตอนได้ดังนี้ [3]

1. วิเคราะห์แหล่งข้อมูล เช่น ปริมาณของข้อมูล จำนวนและชนิดของการเข้าถึงแหล่งข้อมูล แพลตฟอร์มและภาษาโปรแกรมที่ใช้ เป็นต้น

2. ย้ายข้อมูลที่ต้องการจากระบบเดิมมาไว้ในบริเวณที่ใช้ปรับแต่งข้อมูล หรือเรียกบริเวณนี้ว่า staging area เพื่อนำมาเลือกเฉพาะส่วนที่ต้องการแปลงข้อมูลและตรวจสอบความถูกต้อง หรือการทำความสะอาดข้อมูล

3. กำหนด primary key ของ fact table และ dimension table และกำหนด foreign key ระหว่าง fact table กับ dimension table

4. ย้ายข้อมูลที่ทำความสะอาดแล้วจาก staging area ลงสู่เซอร์ฟเวอร์ของดาต้ามาร์ท

5. สร้าง metadata ของแต่ละดาต้ามาร์ท โดยเก็บรายละเอียดของข้อมูลการอัพเดตและส่งออกไปไว้ในดาต้ามาร์ท

6. ตรวจสอบความถูกต้องของข้อมูล ซึ่งจะต้องกระทำตลอดทั้งกระบวนการแปลงข้อมูล จะทำได้ดังนี้

(1) ตรวจสอบผลรวมทั้งหมดของจำนวนข้อมูลที่ดึงมาจากแหล่งข้อมูลกับข้อมูลที่เพิ่มเข้าไป

(2) ตรวจแก้ข้อมูลในระบบเดิมของแหล่งข้อมูล หรือในรูทีนของการแปลง ซึ่งควรจะเก็บข้อมูลในการตรวจแก้ไว้ใน metadata ของการแปลงข้อมูลด้วย

(3) ตรวจสอบค่าของข้อมูลให้ถูกต้องในกระบวนการรวบรวมข้อมูล

(4) ตรวจสอบผลรวมของข้อมูลหลังจากการย้ายข้อมูลลงสู่ดาต้ามาร์ทแล้ว

4.4 สิ่งที่ควรพิจารณาก่อนสร้างคลังข้อมูล

เนื่องจากการลงทุนสร้างคลังข้อมูลขึ้นมาใช้เพื่อสนับสนุนการทำงานขององค์กรนั้น จำเป็นต้องมีค่าใช้จ่ายในการลงทุนมหาศาล ทั้งที่สามารถวัดออกมาเป็นตัวเงินได้ เช่น ค่าใช้จ่ายด้านฮาร์ดแวร์ ซอฟต์แวร์ และ infrastructure อื่น ๆ ที่จำเป็นต้องใช้ ส่วนค่าใช้จ่ายที่ไม่เป็นตัวเงินแต่มีความสำคัญ

อย่างมาก ได้แก่ กำลังแรงงานที่จะเสียไปของทรัพยากรบุคคลขององค์กร และเวลาที่ใช้ไปกับการพัฒนา ดังนั้น เมื่อองค์กรตัดสินใจสร้างคลังข้อมูลขึ้นแล้ว ควรจะประสบความสำเร็จด้วย ทั้งนี้ Poe [6] ได้เสนอ The Big Eight หรือ 8 ประการที่ควรให้ความสนใจ โดยมีรายละเอียดดังนี้

1. ควรมีความชัดเจนในเป้าหมายร่วมของการสร้างระบบนี้ของคนในองค์กร เหมือนการตอบคำถามว่า ทำไมคุณถึงคิดจะสร้างคลังข้อมูล? ซึ่งคำตอบขององค์กรที่จะได้ คือ เป้าหมายที่ต้องการ โดยควรจะเขียนเป้าหมายนี้ออกมาเป็นลายลักษณ์อักษรที่ชัดเจน เพื่อให้ทีมพัฒนาได้เข้าใจเป้าหมายร่วมกัน

2. ทำความเข้าใจสถาปัตยกรรมของระบบ เพื่อให้ทีมพัฒนาเข้าใจตรงกัน ในที่นี้หมายถึง blueprint ที่แสดง E-R model รวมของระบบ ความเข้าใจที่ตรงกันทำให้งานเดินไปได้เร็วขึ้น

3. เทคโนโลยีที่ใช้ควรอยู่ในวิสัยที่เหมาะสม ทั้งด้านของตัวเงินและความยากง่ายในการเรียนรู้ ทั้งนี้หมายรวมทั้ง ฮาร์ดแวร์ ซอฟต์แวร์ และเครือข่าย อาจต้องมีการทดสอบและฝึกอบรมก่อนการใช้งานจริง

4. ทีมพัฒนาต้องมีวิสัยทัศน์เชิงบวกในการทำงาน เนื่องจากทีมพัฒนามักมาจากส่วนงานด้านเทคโนโลยีสารสนเทศ แต่ในเนื้องานจริง ๆ แล้วผู้ใช้ขั้นปลายเป็นส่วนงานอื่น ๆ ขององค์กร ดังนั้น จึงจำเป็นอย่างยิ่งที่จะให้ผู้ใช้ขั้นปลายที่เป็นเจ้าของงานเข้ามาร่วมทำงานด้วยตั้งแต่ต้นโครงการ

5. ต้องมั่นใจได้ว่าทีมพัฒนาเข้าใจเป็นอย่างดีถึงความแตกต่างกันระหว่างฐานข้อมูลปฏิบัติการและฐานข้อมูลสนับสนุนการตัดสินใจ

6. จัดให้มีการฝึกอบรม โดยควรเป็นการฝึกอบรมก่อนเริ่มโครงการ โดยเฉพาะอย่างยิ่งการฝึกอบรมเกี่ยวกับเครื่องมือที่องค์กรจะใช้ในการพัฒนา ทั้งนี้อาจเป็นการฝึกอบรมจากบริษัทผู้ขาย

7. ควรหาบุคลากรที่มีประสบการณ์ในการพัฒนาคลังข้อมูล เพื่อทำหน้าที่เป็นผู้จัดการโครงการ หรือถ้าในองค์กรไม่เคยมีประสบการณ์เลย อาจจ้างที่ปรึกษาที่มีความเชี่ยวชาญและมีประสบการณ์ด้านนี้โดยเฉพาะมาช่วยทีมพัฒนา

8. โปรแกรมที่จะใช้นำเสนอข้อมูลในคลังข้อมูลต้องสามารถเรียนรู้ได้ง่าย และผู้ใช้สามารถใช้งานได้อย่างมีประสิทธิภาพ

5. สรุป

คลังข้อมูลเป็นการรวบรวมข้อมูลจากฐานข้อมูลของระบบงานปฏิบัติงานประจำวันขององค์กร แล้วนำมาแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมในการเก็บและสะดวกในการใช้งาน แล้วจึงนำข้อมูลนั้นเข้าไปเก็บในคลังข้อมูล

การพัฒนาหรือสร้างคลังข้อมูลมาใช้ในองค์กรจะต้องมีการพิจารณาถึงองค์ประกอบที่จำเป็นในการสร้างให้เหมาะสมด้วย ทั้งนี้เพื่อให้เกิดความคุ้มค่าในการลงทุนและเกิดประโยชน์สูงสุดต่อองค์กร ถึงแม้ว่าเทคโนโลยีคลังข้อมูลจะให้ประสิทธิภาพในการใช้ข้อมูลอย่างมากก็ตาม แต่สิ่งที่ต้องคำนึงถึงด้วยคือ ทรัพยากรที่องค์กรจะต้องทุ่มเทลงไปในการพัฒนาระบบนี้ มีทั้งที่สามารถวัดเป็นตัวเงินได้และที่ไม่สามารถตีค่าออกมาเป็นตัวเงินได้ นอกจากนี้ ปัญหาในระหว่างการพัฒนาที่อาจจะเกิดขึ้น จนองค์กรไม่สามารถจะพัฒนาระบบนี้จนสำเร็จ และนำมาใช้งานได้ เกิดการลงทุนที่สูญเปล่า ดังนั้นจึงต้องมีการวางแผนควบคุมและจัดการให้รอบคอบ

เอกสารอ้างอิง

[1] กองบรรณาธิการ, “Data Warehouse เปิดผนึกคลังข้อมูลอัจฉริยะ”, *วารสาร IT Sof* , vol 6, no.64, หน้า 120, กรกฎาคม 2540.

[2] คมกริช ศิริแสงชัยกุล, “Data Warehouse ระบบการจัดการไอที”, *สารเนคเทค*, ปีที่ 7, ฉบับที่ 31, หน้า 37, พฤศจิกายน - ธันวาคม 2542.

[3] นงลักษณ์ พลอยปลื้ม, “การแปลงข้อมูลเข้าสู่ Data Warehouse”, *BCM Magazine*, vol 8, no.103, หน้า 126, กันยายน 2540.

[4] เลิศ เลิศศิริโสภณ, “ถึงเวลาของดาต้าแวร์เฮ้าส์แล้วหรือยัง”, *BCM Magazine*, vol 9, no.115, หน้า 95, กันยายน 2541.

[5] Mattison, Rob, *Data Warehousing*, New York, NY : McGraw-Hill , 1996.

[6] Poe, Vidette, *Buliding a Data Warehouse for Decision Support*, Upper Saddle River, NJ : Prentice Hall, 1996.

[7] Connolly, Thomas and Begg, Carolyn, *Database Systems: A Practical Approach to Design, Implementation, and Management*, Reading, MA : Addison Wesley, 2002.

[8] Pantip. *Data Warehouse* . [Online]. Available : http://www.pantip.com/magazine/feature/apr_datawarehouse.htm . 1991.