บทที่ 2 การทำความสะอาดข้อมูล (DATA CLEANSING)

ผู้ช่วยศาสตราจารย์จุฑาวุฒิ จันทรมาลี

หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสวนดุสิต

1. DATA CLEANSING

หมายถึง กระบวนการตรวจสอบ สะสาง แก้ไข หรือจัดรูปแบบข้อมูลให้อยู่ในสภาพที่พร้อมใช้งานที่สุด รวมไปถึงคัดกรองข้อมูลที่ไม่ถูกต้องหรือไม่จำเป็นออกไปจากชุดข้อมูลที่จะใช้วิเคราะห์หรือประมวลผล เพื่อให้ชุดข้อมูลที่จะใช้มีความสมบูรณ์ มีคุณภาพ พร้อมนำไปวิเคราะห์และใช้ประโยชน์ ซึ่งอาจเรียกอีกอย่างว่า เป็นการทำ “re-oganize” ข้อมูลใหม่ก็ได้

เหตุผลที่ต้องทำ Data Cleansing

1. เพื่อกำจัดข้อผิดพลาดของข้อมูลที่มาจากหลากหลายฐานข้อมูล
2. ช่วยให้จับคู่ชุดข้อมูลกับการวิเคราะห์เพื่อหาคำตอบได้ง่ายขึ้น เร็วขึ้น
3. ช่วยให้เห็นรูปแบบของข้อมูลที่ผิดพลาดเพื่อป้องกันและปรับปรุงการนำเข้าข้อมูล เช่น การกรอกแบบฟอร์ม การคีย์ข้อมูลลงระบบ
4. ช่วยให้ได้ Insight หรือรายงาน (report) ที่แม่นยำ ทำให้ตัดสินใจได้รวดเร็วขึ้น
5. ช่วยให้ดึงข้อมูลออกมาใช้ได้ทันที พร้อมข้อมูลอยู่ในรูปแบบที่สมบูรณ์
6. ช่วยล้างข้อมูลที่หมดอายุ ซึ่งเกี่ยวข้องกับ พ.ร.บ. ข้อมูลส่วนบุคคล (PDPA)

2. ลักษณะข้อมูลที่ต้องทำ DATA CLEANSING

1. ชุดข้อมูลที่ไม่ได้อยู่ในรูปแบบหรือไฟล์ประเภทเดียวกัน

กรณีนี้ อาจเกิดจากการที่มีข้อมูลจากหลากหลายฐานข้อมูล (database) เมื่อมาจากหลากหลายแหล่ง จึงมีโอกาสที่แต่ละแหล่งจะจัดเก็บข้อมูลกันคนละสกุลไฟล์ เช่น จัดเก็บข้อมูลเป็นไฟล์ .pdf ไฟล์ excel หรือสกุลไฟล์อื่น ๆ ที่ไม่คุ้นเคย เมื่อจำเป็นต้องนำมาประมวลผลด้วยกัน จึงไม่สามารถทำได้

การทำ Data Cleansing สำหรับข้อมูลประเภทนี้ จึงเป็นการแปลงไฟล์ข้อมูลต่างๆ ให้เป็นไฟล์สกุลเดียวกัน เพื่อให้สามารถนำมาประมวลผลด้วยกันได้ นอกจากนี้ การจัดรูปแบบไฟล์ใหม่ อาจทำไปเพื่อลดการใช้พื้นที่จัดเก็บ เช่น การรวมไฟล์ข้อมูลเข้าไปไฟล์เดียว การแปลงข้อมูลมาอยู่ในรูปแบบไฟล์ที่กินพื้นที่น้อยกว่า

2. ข้อมูลไม่ได้ถูกจัดเก็บในรูปแบบที่ต้องการ

จำเป็นต้องทำให้ข้อมูลอยู่ในรูปแบบที่พร้อมสำหรับการวิเคราะห์ข้อมูลหรือนำไปสู่ report หรือคำตอบที่เราต้องการได้ ยกตัวอย่างเช่น ต้องการทราบร้านคู่แข่งในแต่ละเขต/อำเภอ แต่ข้อมูลที่มีอยู่นำเข้ามาจาก Open Source ที่แบ่งพื้นที่ตามพิกัด GPS (ละติจูด-ลองติจูด) เราจึงต้องแปลงข้อมูลที่ได้ให้เป็นแบบเขต/อำเภอ ก่อน

หรืออีกตัวอย่างง่ายๆ คือ ข้อมูลไม่ได้ถูกเก็บอยู่ในไฟล์ที่สามารถส่งข้อมูลเข้าประมวลผลได้ เช่น ได้ข้อมูลเป็นไฟล์รูปภาพ (.jpg หรือ .png) อาจต้องแปลงเป็นไฟล์ข้อความหรือสคริปต์ (.csv, .tsv, .json, .xml) เพื่อให้พร้อมสำหรับการวิเคราะห์ก่อน

3. ข้อมูลที่ไม่ถูกต้องหรือมีข้อผิดพลาดในการเก็บข้อมูล

ส่วนใหญ่ข้อมูลที่กรอกหรือนำเข้าสู่ระบบโดยคน อาจมีข้อผิดพลาด (human errors) บ้างเป็นธรรมดา เช่น การกรอกข้อมูลผิดช่อง การกรอกข้อมูลที่ไม่มีทางเป็นความจริง การกรอกข้อมูลที่ถูกต้องแต่ไม่ตรงกับข้อมูลหลัก (ตัวอย่างชื่อร้านเดียวกัน แต่เขียนหรือสะกดต่างกัน) ฯลฯ ทำให้ได้ชุดข้อมูลที่เมื่อประมวลผลออกมาแล้วได้ Insight หรือคำตอบที่ไม่แม่นยำหรือผิดจากความจริงไป

นอกจากข้อผิดพลาดที่เกิดขึ้นจากคนแล้ว ยังอาจมีความคลาดเคลื่อนของระบบได้ เช่น การติดตามกิจกรรมต่างๆ (event tracking) ที่เกิดขึ้นในระบบ ไม่ได้ถูกบันทึกตามเวลาจริง เพราะอาจมีความล้าช้าจากการส่งข้อมูล หรือเกิดความขัดข้องอื่นๆ

3. กระบวนการทำ DATA CLEANSING

1. กำจัดข้อมูลที่ซ้ำซ้อนและข้อมูลที่ไม่เกี่ยวข้องออก

คือ การกำจัดข้อมูลที่ซ้ำซ้อนและข้อมูลที่ไม่เกี่ยวข้องเพื่อลดภาระในการประมวลผลและให้ได้ข้อมูลที่ 'เกลี้ยงเกลา' ที่สุด

1.1 ข้อมูลที่ซ้ำซ้อน (duplicated data) มักจะได้มาจากการกรอกข้อมูลซ้ำซ้อนของคน การดึงข้อมูลจากหลายแหล่ง หลายแผนก หรือการรวมข้อมูลภายในและข้อมูลจาก Open source เพราะมีโอกาสสูงที่แหล่งข้อมูลแต่ละแหล่งจะขอ/คีย์ข้อมูลชุดเดียวกัน หากไม่นำข้อมูลที่ซ้ำซ้อนออกก่อนที่จะวิเคราะห์ข้อมูล Insight ที่ได้อาจจะไม่ตรงกับความเป็นจริง

1.2 ข้อมูลที่ไม่เกี่ยวข้อง (Irrelevant data) นั้น ส่วนมากจะเกิดขึ้น จากคำตอบหรือเป้าหมายในการวิเคราะห์ข้อมูลที่เราต้องการที่แตกต่างกันออกไป ทำให้ไม่มีความจำเป็นต้องใช้ข้อมูลบางชุด เช่น เรามีข้อมูลกิจกรรมการขายทุกอย่าง หากต้องการรู้เฉพาะว่า สินค้าใดขายดี ก็ไม่จำเป็นต้องนำข้อมูลกลุ่มลูกค้าที่ซื้อมาคำนวณ ดังนั้นก่อนที่จะทำการวิเคราะห์ประมวลผล จึงควรนำข้อมูลที่ไม่จำเป็น ไม่ใช่ปัจจัยที่จะทำให้ได้คำตอบออกไปก่อน

2. แก้ไขข้อผิดพลาดในเชิงโครงสร้างหรือรูปแบบ

คือ การจัดการกับข้อมูลที่ผิดพลาดในเชิงโครงสร้างหรือรูปแบบ (structural errors) เช่น ข้อมูลข้อมูลหนึ่งหรือคำคำหนึ่งที่มีค่าเท่ากัน ระบบอาจมองว่าเป็นข้อมูลคนละตัว การทำ Data Cleansing ในขั้นตอนนี้คือการกำหนดให้ข้อมูลที่หมายถึงสิ่งเดียวกันมีค่าเท่ากัน ยกตัวอย่างเช่น กำหนด “N/A” กับ “Not Applicable” ให้มีค่าเท่ากัน รวมไปถึงการกำหนดการคำนวณทศนิยม ว่าต้องการเอากี่หลักหรือต้องการปัดขึ้นหรือลง

3. กรองข้อมูลที่มีค่าผิดปกติออกจากชุดข้อมูล

บางทีข้อมูลที่ได้ โดยเฉพาะข้อมูลที่นำเข้าโดยมนุษย์ เช่น กรอกข้อมูล พิมพ์ ๆลๆ มีโอกาสที่จะได้ค่าที่ไม่มีทางเป็นไปได้ ยกตัวอย่างเช่น จำนวนสมาชิกในครอบครัว 500 คน ซึ่งระบบไม่รู้ว่าเป็นข้อมูลที่เกินจริง หรืออัตราการซื้อสินค้าที่เกินสินค้าในคลัง หรือตัวเลขอื่นๆ ที่น้อยหรือมากกว่าที่ควรจะเป็น การทำ Data Cleansing ในขั้นตอนนี้ คือ การกำจัดช่วงของค่าหรือตัวเลขที่ไม่มีทางเป็นจริงออกในคราวเดียว ด้วยการกำหนด “Outliner” ขีดเส้นกั้นเอาเฉพาะช่วงข้อมูลที่ต้องการ

4. จัดการกับข้อมูลที่หายไปหรือไม่สมบูรณ์

หากมีข้อมูลที่หายไป ผลลัพธ์หรือคำตอบจากการวิเคราะห์ย่อมเปลี่ยนไป ความแม่นยำลดลง แต่ยิ่งไปกว่านั้น หากมีข้อมูลที่หายไป อัลกอริทึม (algorithm) หรือระบบที่ใช้วิเคราะห์อาจไม่ทำงานได้ ดังนั้น เราจึงต้องจัดการกับข้อมูลที่หายไปก่อนที่จะทำการประมวลผลข้อมูล โดยวิธีที่อาจจะทำได้ก็อย่างเช่น

4.1 ตัดข้อมูลที่หายไปออกจากการประมวลผล ซึ่งหากใช้วิธีนี้ อาจจะต้องใคร่ครวญดีๆ ก่อนว่าคุ้มค่าที่จะทำหรือไม่ เพราะจะทำให้ปัจจัยในการประมวลผลหายไปหนึ่งปัจจัย คำตอบที่ได้ก็จะไม่ละเอียดเท่าตอนที่มีปัจจัยให้วิเคราะห์ครบถ้วน

4.2 ใส่หรือแทนค่าข้อมูลที่หายไป ซึ่งอาจได้จากการไปหาข้อมูลเพิ่มหรือจากการสันนิษฐาน อย่างไรก็ตาม หากข้อมูลหายไปเป็นจำนวนมาก อาจต้องใช้เวลาในการใส่ข้อมูลใหม่

4.3 เปลี่ยนวิธีประมวลผลหรือใช้ชุดข้อมูลอื่นๆ ที่สามารถให้ Insight ได้ใกล้เคียง

5. ตรวจสอบความถูกต้อง

คือ ขั้นตอนการตรวจสอบชุดข้อมูลว่าได้ชุดข้อมูลที่สมบูรณ์พร้อมสำหรับการวิเคราะห์เพื่อให้ได้คำตอบที่ต้องการแล้วหรือไม่ โดยเราอาจตรวจสอบได้ง่ายๆ ด้วยการถามคำถามกับตัวเอง เช่น

5.1 ชุดข้อมูลที่ได้มีความสมเหตุสมผลหรือไม่

5.2 ข้อมูลอยู่ในรูปแบบที่เหมาะสมสำหรับการหาคำตอบหรือไม่

5.3 ชุดข้อมูลที่ได้สามารถให้คำตอบหรือ Insight ที่ต้องการได้จริงหรือไม่

หากไม่แน่ใจ ว่าชุดข้อมูลที่ได้สามารถให้คำตอบได้ อาจต้องทำความสะอาดข้อมูลอีกรอบ ขิด หรือเลือกใช้ชุดข้อมูลชุดใหม่ในการวิเคราะห์

4. ข้อควรระวังในการทำ DATA CLEANSING

1. การพิมพ์ผิด เมื่อมีการลงรายการใด ๆ ก็ตามในฐานข้อมูล ควรจะมีการตรวจสอบอย่างละเอียด รอบคอบและถี่ถ้วน และถี่ถ้วน
2. การลงรายการที่ไม่ครบถ้วน ขาดรายการในบางเขตข้อมูล โดยเฉพาะเขตข้อมูลที่จำเป็น
3. การไม่สม่ำเสมอในการลงรายการของข้อมูล ด้วยคำ ๆ เดียวกัน แต่มีการใช้ไม่เหมือนกัน แล้วแต่ความสะดวก ไม่มีมาตรฐานในการลงรายการ
4. การตรวจสอบไม่ละเอียด ทำให้เกิดรายการซ้ำขึ้นได้ และทำให้เกิดการประมวลผลผิดพลาด
5. การไม่ปรับปรุงรายการหลักฐานให้ทันสมัยอยู่เสมอ

#	ID	Name	LastName	Birthday	Join date	Gender	Tel	E-mail
1	1111	ธนวัฒน์	เจริญด้วยทรัพย์	1-1-1987	01/01/2019	M	860147805	username @ riccoprint.com
2	1112	วัชสัณห์	เจริญด้วยจิตใจ	07/07/1987	01/01/2019	M	087-0147-805	username2@riccoprint.com
3	1113	พลลภัตม์	กำไลอันประเสริฐ	1987/07/07	01/01/2019	M	0810147808	username.riccoprint.com
4	1114	มนวรรธน์		05/07/2530	01/01/2019	F	0820147805	username3@ riccoprint.com
5	1115	ขนมพันธ์	ผูกพันกับหนังสือ	09/12/1988	01/01/2019	A	0830147805	username4@riccoprint.com

Missing values Formats Invalid values Formats Formats

5. DATA CLEANSING TOOLS

1. Integrate.io คือ เครื่องมือจัดการข้อมูลระดับสูง มีฟังก์ชันการทำ ETL, ELT ที่ผู้ใช้งานสามารถตั้งค่าฟังก์ชันต่าง ๆ ด้วยอินเตอร์เฟซที่ใช้งานง่าย ไม่ต้องใช้โค้ด ช่วยทำความสะอาดและแปลงข้อมูลก่อนส่งไปยังแหล่งเก็บข้อมูลต่าง ๆ ไม่ว่าจะเป็น Data Lake, Database หรือ Salesforce ทำให้ Integrate.io เป็นหนึ่งใน Data Cleansing Tools ที่ใช้กันอย่างแพร่หลาย

ประโยชน์ของ Integrate.io

1.1 มีอินเตอร์เฟซที่เป็นมิตรกับผู้ใช้และไม่ต้องใช้โค้ด

1.2 ทำความสะอาดและแก้ไขข้อมูลก่อนส่งไปยังคลังข้อมูล

1.3 เป็นแพลตฟอร์มบนคลาวด์

2. Tibco Clarity คือ เครื่องมือ Data Cleansing เชิงโต้ตอบ ที่ใช้อินเตอร์เฟซแบบภาพเพื่อช่วยให้การปรับปรุงคุณภาพข้อมูล การค้นหาข้อมูล และการแปลงข้อมูลมีประสิทธิภาพมากขึ้น สามารถรองรับข้อมูลดิบทุกประเภท กำจัดข้อมูลที่มีความซ้ำซ้อน และตรวจสอบก่อนที่จะเคลื่อนย้ายข้อมูลไปยังปลายทาง นอกจากนี้ Tibco Clarity ยังสามารถแสดงผลข้อมูลในรูปแบบต่าง ๆ ซึ่งช่วยให้เข้าใจข้อมูลชุดนั้นได้ดีขึ้น

ประโยชน์ของ Tibco Clarity

2.1 มีอินเตอร์เฟซ Data Cleansing แบบภาพ ช่วยให้เข้าใจง่ายขึ้น

2.2 แสดงผลข้อมูลในรูปแบบต่าง ๆ (Data visualizations)

2.3 สามารถตรวจสอบข้อมูลได้ตามกฎที่กำหนด

3. WinPure Clean & Match คือ เครื่องมือ Data Cleansing ที่นิยมใช้กันอย่างแพร่หลาย ช่วยทำความสะอาด ลบข้อมูลที่ซ้ำซ้อน และแก้ไขข้อมูล เหมาะสำหรับข้อมูลธุรกิจและข้อมูลผู้บริโภคที่อยู่ใน Data Base, CRM Data และ Spreadsheets นอกจากนี้ WinPure Clean & Match ยังเป็นเครื่องมือที่ใช้งานง่าย จึงเหมาะสำหรับผู้ใช้ที่ไม่เชี่ยวชาญด้านเทคนิค หรือธุรกิจขนาดเล็กที่มีทรัพยากรด้าน IT จำกัดนั่นเอง


QUESTIONS
ANSWERS