Lab 1: จำแนกประเภทและออกแบบการจัดเก็บข้อมูล

ตารางวิเคราะห์ประเภทข้อมูลและแหล่งจัดเก็บ

ที่	ข้อมูล	ประเภทข้อมูล	รูปแบบไฟล์ / Schema	แหล่งจัดเก็บที่เหมาะสม
1	ตารางข้อมูลนิสิต (CSV)	Structured Data	ไฟล์ CSV มีคอลัมน์คงที่ (student_id, name, faculty, year) แต่ละแถวคือ 1 เรคอร์ด มี Schema ชัดเจนตายตัว	Relational Database เช่น MySQL/PostgreSQL)เพื่อรองรับการ Query, JOIN กับตารางอื่น และควบคุมความถูกต้องของข้อมูล
2	ข้อมูลเซนเซอร์อุณหภูมิ (JSON)	Semi-structured Data	รูปแบบ JSON มี key-value และมีโครงสร้างแบบ nested ได้ แต่ Schema ยืดหยุ่น เปลี่ยนแปลง field ได้ตามรุ่นอุปกรณ์	Data Lake เก็บ JSON คิบร่วมกับ Time-series Database เช่น InfluxDB สำหรับการวิเคราะห์แนวโน้มค่าตามเวลา
3	ภาพจากกล้องวงจรปิด	Unstructured Data	ไฟล์ภาพ/วิดีโอ (JPEG, MP4) เป็นข้อมูล Binary ไม่มีโครงสร้างแบบตาราง	Object Storage เช่น Amazon S3, MinIO ที่ออกแบบมาสำหรับไฟล์ขนาดใหญ่จำนวนมาก พร้อมทำ Metadata index แยกต่างหาก
4	ไฟล์ PDF ใบสมัครทุน	Unstructured Data	ไฟล์ PDF ผสมข้อความและรูปแบบฟอร์ม แม้เนื้อหาจะมีโครงสร้างบางส่วนแต่ไฟล์ระดับ Storage ถือเป็น Binary/Unstructured	Object Storage หรือ Document Management System พร้อมทำดัชนี Metadata เลขที่คำร้อง, วันที่ยื่น เก็บใน Relational DB แยก
5	Log การเข้าใช้งานระบบ LMS	Semi-structured Data	ไฟล์ Log เป็นข้อความรายบรรทัด หรือ JSON log มีรูปแบบคงที่บางส่วน เช่น timestamp, user_id, action แต่ไม่ใช่ตารางเชิงสัมพันธ์เต็มรูปแบบ	Data Lake หรือระบบจัดการ Log โดยเฉพาะ เช่น Elasticsearch/ELK Stack เพื่อการค้นหาและวิเคราะห์ Log จำนวนมาก
6	ตารางผลการลงทะเบียนใน MySQL	Structured Data	ตารางเชิงสัมพันธ์ (Relational Table) มี Schema ตายตัว เช่น student_id, course_id, grade, semester	Relational Database (MySQL) เป็นแหล่งดำเนินงานหลัก และอาจ ETL เข้า Data Warehouse เพื่อทำรายงานเชิงวิเคราะห์
7	ไฟล์เสียงจากศูนย์รับแจ้งปัญหา	Unstructured Data	เป็นไฟล์เสียง (WAV, MP3) เป็นข้อมูล Binary ไม่มีโครงสร้างแบบตาราง ต้องผ่าน Speech-to-Text จึงจะดึงข้อความออกมาได้	Object Storage สำหรับเก็บไฟล์เสียงต้นฉบับ ร่วมกับ Data Lake สำหรับเก็บผลลัพธ์ที่ผ่านการแปลงเป็นข้อความแล้ว
8	ข้อมูลดิบจากเครื่องอ่านบัตรที่ยังไม่ตรวจสอบ	Raw Data	ข้อมูลดิบที่ยังไม่ได้ทำความสะอาดหรือตรวจสอบความถูกต้อง อาจอยู่ในรูปแบบ CSV/Log ที่มี Error หรือค่าซ้ำปะปน	Data Lake (Raw / Landing Zone) เพื่อเก็บข้อมูลต้นฉบับไว้ก่อน แล้วจึงทำ ETL/Cleaning ไปเก็บใน Data Warehouse ต่อไป

สรุปหลักการจำแนกประเภทข้อมูล

- Structured Data: ข้อมูลมี Schema ตายตัว จัดเก็บเป็นตาราง แถว/คอลัมน์ชัดเจน เช่น ข้อมูลนิสิต และผลการลงทะเบียน
- Semi-structured Data: มีโครงสร้างบางส่วน (เช่น key-value, tag) แต่ Schema ยืดหยุ่นเปลี่ยนแปลงได้ เช่น JSON จากเซนเซอร์ และ Log ของ LMS
- Unstructured Data: ไม่มีโครงสร้างตารางแบบตายตัว เช่น ภาพ วิดีโอ ไฟล์เสียง และเอกสาร PDF
- Raw Data: ข้อมูลดิบที่ยังไม่ผ่านการตรวจสอบ/ทำความสะอาด อาจมี Error หรือรูปแบบไม่แน่นอนปะปนอยู่ ต้องผ่านกระบวนการตรวจสอบก่อนนำไปใช้งานจริง