1.1 ที่มาของเรื่อง

ในปัจจุบันธุรกิจพาณิชย์อิเล็กทรอนิกส์ (E-Commerce) มีการเติบโตอย่างรวดเร็วและต่อเนื่อง ส่งผลให้เกิดข้อมูลจำนวนมหาศาลจากการทำธุรกรรมในแต่ละวัน ไม่ว่าจะเป็นข้อมูลคำสั่งซื้อ ข้อมูลสินค้า ข้อมูลลูกค้า ข้อมูลผู้ขาย ข้อมูลการชำระเงิน ตลอดจนความคิดเห็นและคะแนนรีวิวจากผู้ซื้อ ข้อมูลเหล่านี้ล้วนมีคุณค่าและสามารถนำมาวิเคราะห์เพื่อสนับสนุนการตัดสินใจเชิงธุรกิจได้

อย่างไรก็ตาม ข้อมูลดังกล่าวมักถูกจัดเก็บกระจัดกระจายอยู่ในหลายตารางและหลายระบบ ซึ่งถูกออกแบบมาเพื่อรองรับการทำธุรกรรม (Transactional System) เป็นหลัก โดยเน้นความถูกต้องของข้อมูลและความเร็วในการบันทึกรายการ จึงไม่เหมาะกับการนำมาวิเคราะห์เชิงลึก เนื่องจากการเชื่อมโยงข้อมูลจำนวนมากระหว่างตารางทำได้ช้า และยังส่งผลกระทบต่อประสิทธิภาพของระบบงานหลักที่ต้องให้บริการผู้ใช้อยู่ตลอดเวลา

ด้วยเหตุนี้ แนวคิดเรื่องคลังข้อมูล (Data Warehouse) จึงถูกนำมาใช้เพื่อรวบรวมข้อมูลจากหลายแหล่งมาจัดเก็บในรูปแบบที่เหมาะสมต่อการวิเคราะห์ โดยแยกออกจากระบบงานหลักอย่างชัดเจน อย่างไรก็ตาม การจัดตั้งคลังข้อมูลด้วยเครื่องแม่ข่ายภายในองค์กรจำเป็นต้องใช้เงินลงทุนสูงในการจัดหาฮาร์ดแวร์ และต้องดูแลรักษาอย่างต่อเนื่อง ซึ่งเป็นข้อจำกัดสำคัญสำหรับองค์กรขนาดเล็กและผู้ที่เริ่มต้นศึกษา

นอกจากนี้ ในสภาพแวดล้อมการทำงานจริง ข้อมูลไม่ได้ถูกนำเข้าสู่ระบบเพียงครั้งเดียว แต่จะทยอยเข้ามาอย่างต่อเนื่องตามรอบเวลาของธุรกิจ การออกแบบระบบคลังข้อมูลจึงควรรองรับการนำเข้าข้อมูลเพิ่มเติมได้โดยอัตโนมัติ โดยไม่ต้องอาศัยการสั่งงานด้วยมือทุกครั้ง เพื่อลดภาระของผู้ดูแลระบบและลดความผิดพลาดที่อาจเกิดขึ้น

ผู้จัดทำจึงมีแนวคิดในการพัฒนาระบบคลังข้อมูลบนบริการคลาวด์ของ Amazon Web Services (AWS) โดยนำชุดข้อมูลจริงจากธุรกิจอีคอมเมิร์ซ Olist ประเทศบราซิล ซึ่งประกอบด้วยคำสั่งซื้อกว่า 100,000 รายการ ระหว่างปี ค.ศ. 2016 ถึง 2018 มาผ่านกระบวนการนำเข้า แปลง และจัดเก็บในรูปแบบ Star Schema พร้อมทั้งออกแบบให้กระบวนการทำงานเริ่มต้นได้เองเมื่อมีข้อมูลชุดใหม่เข้าสู่ระบบ เพื่อแสดงให้เห็นว่าการสร้างคลังข้อมูลสามารถทำได้โดยไม่ต้องลงทุนกับฮาร์ดแวร์จริง อีกทั้งยังสามารถปรับขยายทรัพยากรได้ตามความต้องการ และคิดค่าใช้จ่ายตามการใช้งานจริง ซึ่งเป็นการใช้ทรัพยากรอย่างคุ้มค่าและยั่งยืน

1.2 วัตถุประสงค์

โครงงานนี้จัดทำขึ้นโดยมีวัตถุประสงค์หลักดังต่อไปนี้

1. เพื่อศึกษาและออกแบบสถาปัตยกรรมคลังข้อมูล (Data Warehouse) บนระบบคลาวด์ โดยใช้บริการของ AWS
2. เพื่อพัฒนากระบวนการ ETL สำหรับนำเข้าและแปลงข้อมูลจากไฟล์ต้นทางให้อยู่ในรูปแบบ Star Schema ที่เหมาะสมต่อการวิเคราะห์
3. เพื่อศึกษาการทำงานของ ETL Pipeline แบบอัตโนมัติ เมื่อมีข้อมูลชุดใหม่ทยอยเข้าสู่ระบบ ตามลักษณะการใช้งานจริงของธุรกิจ
4. เพื่อแสดงให้เห็นว่าการจัดเก็บและประมวลผลข้อมูลขนาดใหญ่สามารถดำเนินการได้โดยไม่จำเป็นต้องมีฮาร์ดแวร์จริง
5. เพื่อเปรียบเทียบรูปแบบการสืบค้นข้อมูลระหว่าง Amazon Athena ซึ่งสืบค้นข้อมูลจาก Amazon S3 โดยตรง กับ Amazon Redshift ซึ่งเป็นคลังข้อมูลโดยเฉพาะ ทั้งในด้านประสิทธิภาพและรูปแบบค่าใช้จ่าย
6. เพื่อนำเสนอผลการวิเคราะห์ข้อมูลเชิงธุรกิจในรูปแบบ Dashboard ที่เข้าใจง่ายและนำไปใช้ประกอบการตัดสินใจได้

1.3 ขอบเขตของระบบ

1. วัตถุประสงค์ของระบบ

ระบบนี้มีวัตถุประสงค์เพื่อนำเทคโนโลยีและเครื่องมือบนระบบคลาวด์มาใช้จัดการวงจรข้อมูลของธุรกิจอีคอมเมิร์ซ ตั้งแต่การนำเข้าข้อมูลดิบ การทำความสะอาดข้อมูล การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ การจัดเก็บในคลังข้อมูล ไปจนถึงการแสดงผลการวิเคราะห์ โดยมุ่งเน้นที่ความถูกต้องของข้อมูล ความสามารถในการทำงานอัตโนมัติ และการใช้ทรัพยากรอย่างคุ้มค่าภายใต้ข้อจำกัดด้านงบประมาณ

2. ฟังก์ชันหลัก

- **ระบบนำเข้าข้อมูลดิบ :** แบ่งชุดข้อมูลต้นทางออกเป็นชุดย่อยตามช่วงเวลา แล้วทยอยนำเข้าสู่พื้นที่จัดเก็บ Raw Zone บน Amazon S3 เพื่อจำลองลักษณะการเข้ามาของข้อมูลจริงที่เกิดขึ้นอย่างต่อเนื่อง

- **ระบบตรวจจับข้อมูลใหม่และสั่งงานอัตโนมัติ :** ตรวจจับเมื่อมีไฟล์ข้อมูลใหม่เข้าสู่ระบบ แล้วสั่งให้กระบวนการ ETL เริ่มทำงานโดยอัตโนมัติ โดยไม่ต้องสั่งงานด้วยมือ
- **ระบบทำความสะอาดข้อมูล :** ตรวจสอบคุณภาพข้อมูลเบื้องต้น จัดการข้อมูลซ้ำ ค่าว่าง และข้อมูลที่ผิดรูปแบบ ก่อนเข้าสู่ขั้นตอนการแปลงข้อมูล
- **ระบบแปลงข้อมูลสู่ Star Schema :** แปลงข้อมูลจากไฟล์ต้นทางให้อยู่ในรูปแบบ Star Schema ประกอบด้วย Fact Table และ Dimension Table พร้อมคำนวณค่าเพิ่มเติมที่จำเป็นต่อการวิเคราะห์ เช่น ระยะเวลาจัดส่ง
- **ระบบจัดเก็บคลังข้อมูล :** จัดเก็บข้อมูลที่ผ่านการแปลงแล้วไว้ในพื้นที่ Curated Zone บน Amazon S3 และในคลังข้อมูล Amazon Redshift
- **ระบบสืบค้นและเปรียบเทียบประสิทธิภาพ :** สืบค้นข้อมูลผ่าน Amazon Athena และ Amazon Redshift เพื่อเปรียบเทียบผลลัพธ์ทั้งด้านประสิทธิภาพและรูปแบบค่าใช้จ่าย
- **ระบบแสดงผลการวิเคราะห์ :** แสดงผลข้อมูลเชิงธุรกิจในรูปแบบ Dashboard ผ่านเว็บแอปพลิเคชันที่พัฒนาขึ้น

3. ข้อกำหนดด้านเทคนิค

ระบบพัฒนาขึ้นบนสภาพแวดล้อม AWS Academy Learner Lab โดยใช้บริการของ AWS ดังแสดงในตารางที่ 1.1 ร่วมกับเทคโนโลยีภายนอกสำหรับการพัฒนาส่วนแสดงผล

ตารางที่ 1.1 บริการของ AWS ที่ใช้ในโครงงานและข้อจำกัดภายใต้ Learner Lab

บริการของ AWS	หน้าที่ในระบบ	ข้อจำกัดใน Learner Lab
Amazon S3	จัดเก็บไฟล์ข้อมูลดิบ (Raw Zone) และข้อมูลที่ผ่านการแปลงแล้ว (Curated Zone) ทำหน้าที่เป็น Data Lake ของระบบ	ใช้งานได้ตามปกติ ใช้สิทธิ์ผ่าน LabRole
AWS Glue DataBrew	ทำความสะอาดข้อมูล เช่น ลบข้อมูลซ้ำ จัดการค่าว่าง และตรวจสอบคุณภาพข้อมูลเบื้องต้น	ใช้ LabRole ได้ ไม่มีข้อจำกัดพิเศษ
AWS Glue (ETL Job)	แปลงข้อมูลจากไฟล์ต้นทางให้อยู่ในรูปแบบ Star Schema และคำนวณค่าเพิ่มเติม เช่น ระยะเวลาจัดส่ง	Worker type: G.1X หรือ Standard, จำนวน Worker สูงสุด 10, Concurrency สูงสุด 1

AWS Glue Data Catalog	จัดเก็บ Metadata และ Schema ของข้อมูลใน S3 เพื่อให้บริการอื่นเรียกใช้ได้	ใช้งานร่วมกับ Glue Crawler ได้ตามปกติ
Amazon Redshift	ทำหน้าที่เป็นคลังข้อมูล จัดเก็บ Fact Table และ Dimension Table สำหรับการวิเคราะห์	รองรับเฉพาะ ra3.large และคลัสเตอร์ขนาดสูงสุด 2 Instance
Amazon Athena	สืบค้นข้อมูลใน S3 โดยตรงแบบ Serverless เพื่อเปรียบเทียบกับ Redshift	ใช้ LabRole ได้ ไม่มีข้อจำกัดพิเศษ
AWS Lambda	ตรวจจับไฟล์ใหม่ที่เข้ามาใน S3 และสั่งให้ ETL Pipeline ทำงานโดยอัตโนมัติ	รันพร้อมกันได้สูงสุด 10 Execution Environment
Amazon EventBridge	ตั้งเวลาให้ Pipeline ทำงานตามรอบที่กำหนด และเชื่อมเหตุการณ์ระหว่างบริการ	ใช้ LabRole ได้ตามปกติ
Amazon EC2	ติดตั้งและให้บริการเว็บแอปพลิเคชันสำหรับแสดงผล Dashboard	รองรับ nano ถึง large เท่านั้น, รันพร้อมกันไม่เกิน 9 Instance และรวมไม่เกิน 32 vCPU
Amazon CloudWatch	ติดตามการทำงานของ ETL Job และเก็บ Log เพื่อตรวจสอบข้อผิดพลาด	ใช้งานได้ตามปกติ
AWS IAM (LabRole)	กำหนดสิทธิ์ให้บริการต่าง ๆ เข้าถึงทรัพยากรระหว่างกันได้	ไม่สามารถสร้าง User, Group หรือ Role ใหม่ได้ ต้องใช้ LabRole ที่เตรียมไว้

สำหรับเทคโนโลยีอื่นนอกเหนือจากบริการของ AWS ที่นำมาใช้ในโครงงาน มีดังนี้

- ภาษา Python สำหรับพัฒนาสคริปต์ ETL และฟังก์ชันการทำงานอัตโนมัติ
- ภาษา SQL สำหรับสร้างโครงสร้างตารางในคลังข้อมูล และสืบค้นข้อมูลเพื่อการวิเคราะห์
- Next.js และ Node.js สำหรับพัฒนาเว็บแอปพลิเคชันแสดงผล Dashboard
- Git และ GitHub สำหรับการควบคุมเวอร์ชันของซอร์สโค้ด

4. ข้อจำกัดของระบบ

ระบบพัฒนาและทดสอบภายใต้สภาพแวดล้อม AWS Academy Learner Lab ซึ่งมีข้อจำกัดที่ส่งผลต่อการออกแบบระบบ ดังนี้

- การใช้งานบริการทั้งหมดจำกัดอยู่เฉพาะ Region us-east-1 และ us-west-2 เท่านั้น
- AWS Glue จำกัด Worker สูงสุด 10 และ Concurrency สูงสุด 1 จึงไม่สามารถรัน ETL Job พร้อมกันหลายงานได้
- Amazon Redshift รองรับเฉพาะ Instance ประเภท ra3.large และคลัสเตอร์ขนาดสูงสุด 2 Instance
- AWS Lambda จำกัดการทำงานพร้อมกันสูงสุด 10 Execution Environment
- Amazon EC2 รองรับเฉพาะ Instance ขนาด nano ถึง large รันพร้อมกันไม่เกิน 9 Instance และรวมกันไม่เกิน 32 vCPU
- ไม่สามารถสร้าง IAM User, Group หรือ Role ใหม่ได้ ต้องใช้ LabRole ที่ระบบเตรียมไว้เท่านั้น
- ไม่สามารถใช้บริการ Amazon QuickSight ได้ จึงต้องพัฒนาเว็บแอปพลิเคชันสำหรับแสดงผล Dashboard ขึ้นเอง
- บัญชีมีงบประมาณจำกัด หากใช้เกินงบที่กำหนดบัญชีจะถูกปิดและทรัพยากรทั้งหมดจะสูญหาย
- ระบบมุ่งเน้นการวิเคราะห์ข้อมูลย้อนหลังแบบ Batch Processing ไม่ครอบคลุมการประมวลผลข้อมูลแบบทันที Real-time Streaming
- ชุดข้อมูลมีขนาดประมาณ 100,000 รายการ ซึ่งเป็นขนาดที่ไม่ใหญ่มาก ผลการเปรียบเทียบด้านความเร็วในการสืบค้นระหว่าง Athena และ Redshift จึงอาจไม่แตกต่างกันอย่างมีนัยสำคัญ

5. ความต้องการด้านความปลอดภัย

- เข้ารหัสข้อมูลที่จัดเก็บบน Amazon S3 ด้วยการเข้ารหัสฝั่งเซิร์ฟเวอร์ (Server-Side Encryption)
- กำหนด Security Group ของ Amazon Redshift และ Amazon EC2 ให้อนุญาตการเข้าถึงเฉพาะแหล่งที่จำเป็นเท่านั้น
- กำหนดสิทธิ์การเข้าถึงทรัพยากรระหว่างบริการผ่าน LabRole ตามข้อจำกัดของสภาพแวดล้อม Learner Lab

- ชุดข้อมูลต้นทางเป็นข้อมูลที่ผ่านการปกปิดตัวตนของลูกค้าและผู้ขายมาแล้วจากผู้เผยแพร่ จึงไม่มีข้อมูลส่วนบุคคลที่สามารถระบุตัวตนได้ในระบบ

6. ข้อกำหนดด้านการใช้งาน

- การเข้าถึงและจัดการทรัพยากรทั้งหมดดำเนินการผ่าน AWS Management Console หรือ AWS CLI ภายในเซสชันของ Learner Lab
- ข้อมูลรับรอง (Credentials) ของสภาพแวดล้อมจะเปลี่ยนใหม่ทุกครั้งที่เริ่มเซสชันใหม่ ผู้ใช้งานจึงต้องตั้งค่าใหม่ก่อนใช้งานทุกรอบ
- ผู้ใช้งานสามารถเข้าถึงผลการวิเคราะห์ได้ผ่านเว็บแอปพลิเคชัน Dashboard โดยไม่จำเป็นต้องเข้าถึงคลังข้อมูลโดยตรง
- ทรัพยากรที่มีค่าใช้จ่ายต่อเนื่อง เช่น คลัสเตอร์ Redshift และ EC2 ควรถูกปิดเมื่อไม่ได้ใช้งานเพื่อรักษางบประมาณของบัญชี

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ระบบคลังข้อมูลที่สามารถรองรับการวิเคราะห์ข้อมูลคำสั่งซื้อกว่า 100,000 รายการได้อย่างมีประสิทธิภาพ
2. เข้าใจหลักการออกแบบ Star Schema รวมถึงการแยก Fact Table และ Dimension Table ในการทำคลังข้อมูล
3. เข้าใจกระบวนการทำงานของ ETL Pipeline บนระบบคลาวด์ ตั้งแต่การนำเข้าข้อมูลจนถึงการนำไปแสดงผล
4. เข้าใจแนวทางการออกแบบระบบให้ทำงานอัตโนมัติเมื่อมีข้อมูลใหม่เข้าสู่ระบบ ซึ่งสอดคล้องกับการใช้งานจริงในภาคธุรกิจ
5. ได้ข้อมูลเชิงลึกทางธุรกิจ เช่น แนวโน้มยอดขายรายเดือน หมวดสินค้าที่ได้รับความนิยม และความสัมพันธ์ระหว่างระยะเวลาจัดส่งกับคะแนนรีวิวของลูกค้า
6. เข้าใจความแตกต่างระหว่างการสืบค้นข้อมูลแบบ Serverless และการใช้คลังข้อมูลเฉพาะทาง ทั้งในด้านประสิทธิภาพและรูปแบบค่าใช้จ่าย
7. สามารถประเมินค่าใช้จ่ายและเลือกใช้บริการคลาวด์ได้อย่างเหมาะสมกับปริมาณงาน
8. ได้รับประสบการณ์ในการวางแผน ติดตามความก้าวหน้า และแก้ไขปัญหาในการทำงานเป็นทีม

1.5 โครงสร้างของรายงาน

รายงานฉบับนี้มีโครงสร้างของรายงานดังต่อไปนี้

บทที่ 1 กล่าวถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ ขอบเขตของระบบ และประโยชน์ที่คาดว่าจะได้รับ

บทที่ 2 ทฤษฎีที่เกี่ยวข้องและเทคโนโลยีที่นำมาใช้ประกอบการศึกษาและพัฒนาระบบคลังข้อมูล

บทที่ 3 ขั้นตอนการดำเนินงาน การวิเคราะห์และออกแบบระบบ ประกอบด้วยการออกแบบสถาปัตยกรรมคลาวด์ การออกแบบ Star Schema และหลักการทำงานของกระบวนการ ETL

บทที่ 4 ผลการดำเนินงาน แสดงผลการทำงานของแต่ละบริการบน AWS ผลการเปรียบเทียบการสืบค้นข้อมูล และผลการแสดงผลผ่าน Dashboard

บทที่ 5 สรุปผลการดำเนินงานของโครงงาน ปัญหาและอุปสรรคที่พบ ข้อจำกัดของระบบ ประมาณการค่าใช้จ่าย และแนวทางการพัฒนาต่อยอด