บทที่ 1

แนะนำ Data Engineering

วิศวกรรมข้อมูล (Data Engineering) หมายถึงกระบวนการออกแบบ พัฒนา และจัดการระบบโครงสร้างข้อมูลที่จำเป็นต่อการจัดเก็บ ประมวลผล และวิเคราะห์ข้อมูลอย่างมีประสิทธิภาพในองค์กร เป็นการเตรียมข้อมูลให้มีคุณภาพและมีความพร้อมสำหรับการนำไปใช้ในงานวิเคราะห์และการทำงานของระบบปัญญาประดิษฐ์ (AI) หรือการเรียนรู้ของเครื่อง (Machine Learning) รวมถึง วิทยาการข้อมูล (Data Science)

แหล่งอ้างอิง:

Schutt, R., & O'Neil, C. (2013). *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media.

Stonebraker, M., & Hellerstein, J. M. (2005). *What Goes Around Comes Around*. Queue, 3(3), 56-59.

Fundamentals of Data Engineering by Joe Reis and Matt Housley (2022), O'Reilly Media.

กระบวนการหลักใน Data Engineering ประกอบด้วย:

1. การรวบรวมข้อมูล (Data Collection): การดึงข้อมูลจากแหล่งต่าง ๆ เช่น ฐานข้อมูล, APIs และไฟล์

2. การจัดการข้อมูล (Data Management): การจัดเก็บข้อมูลในโครงสร้างที่เหมาะสม เช่น Data Warehouse, Data Lake

3. การประมวลผลข้อมูล (Data Processing): เช่น การทำความสะอาดข้อมูล (Data Cleaning) การแปลงรูปแบบข้อมูล (Data Transformation) และการรวมข้อมูลหรือการบูรณาการข้อมูล (Data Integration)

4. การทำข้อมูลให้เป็นไปตามมาตรฐาน (Data Standardization): การจัดการข้อมูลให้เป็นไปตามมาตรฐานที่กำหนด เพื่อให้สามารถนำไปใช้ได้อย่างมีประสิทธิภาพ

5. การตรวจสอบและรักษาความปลอดภัยของข้อมูล (Data Governance and Security): การตรวจสอบและรักษาความปลอดภัยของข้อมูล เพื่อให้แน่ใจว่าข้อมูลถูกต้องและปลอดภัย

เพิ่มเติม บทที่ 1 ภาพรวมวิศวกรรมซอฟต์แวร์ ม.เชียงใหม่

วิทยาการข้อมูล (Data Science)

วิทยาการข้อมูล (Data Science) เป็นสาขาวิชาที่ผสมผสานความรู้ทางคอมพิวเตอร์ สถิติ และความเข้าใจในธุรกิจ เพื่อวิเคราะห์และตีความข้อมูลขนาดใหญ่ (Big Data) โดยมีวัตถุประสงค์หลักในการค้นหารูปแบบและแนวโน้มในข้อมูล ซึ่งนำไปสู่การตัดสินใจที่มีเหตุผลและมีข้อมูลรองรับ

องค์ประกอบหลักของวิทยาการข้อมูล:

การเก็บรวบรวมข้อมูล (Data Collection): การรวบรวมข้อมูลจากแหล่งข้อมูลที่หลากหลาย เช่น ฐานข้อมูล, APIs, เว็บไซต์ และเซนเซอร์ต่าง ๆ

การจัดเก็บข้อมูล (Data Storage): การจัดเก็บข้อมูลในโครงสร้างที่เหมาะสม เช่น ฐานข้อมูลเชิงสัมพันธ์ (Relational Databases), Data Warehouses, และ Data Lakes

การทำความสะอาดข้อมูล (Data Cleaning): การจัดการข้อมูลที่ไม่สมบูรณ์ หรือข้อมูลที่มีความผิดพลาดให้เป็นข้อมูลที่มีคุณภาพและพร้อมใช้งาน

การวิเคราะห์ข้อมูล (Data Analysis): การใช้เทคนิคทางสถิติและการทำเหมืองข้อมูล (Data Mining) เพื่อค้นหารูปแบบและแนวโน้มในข้อมูล

การสร้างแบบจำลอง (Modeling): การสร้างและทดสอบโมเดลทางสถิติและ Machine Learning เพื่อทำนายและวิเคราะห์ข้อมูลในอนาคต

การแสดงผลข้อมูล (Data Visualization): การสร้างกราฟและภาพแสดงข้อมูลเพื่อให้เข้าใจข้อมูลได้ง่ายขึ้น

การสื่อสารและการตัดสินใจ (Communication and Decision Making): การนำเสนอผลการวิเคราะห์ให้กับผู้มีส่วนได้ส่วนเสียเพื่อสนับสนุนการตัดสินใจ

หมายเหตุ ปัจจุบันบางงาน อยู่ใน Data Engineering

ความสำคัญของวิทยาการข้อมูล:

การตัดสินใจที่มีข้อมูลรองรับ: วิทยาการข้อมูลช่วยให้ผู้บริหารและผู้ตัดสินใจสามารถทำการตัดสินใจที่มีเหตุผลและมีข้อมูลรองรับ

การปรับปรุงประสิทธิภาพ: การวิเคราะห์ข้อมูลช่วยในการค้นหาจุดอ่อนและโอกาสในการปรับปรุงกระบวนการทำงานและการดำเนินธุรกิจ

การพัฒนาผลิตภัณฑ์และบริการใหม่: การทำความเข้าใจข้อมูลของลูกค้าและตลาด ช่วยในการพัฒนาผลิตภัณฑ์และบริการที่ตอบสนองความต้องการของลูกค้าได้ดียิ่งขึ้น

Big Data

Big Data หมายถึงชุดข้อมูลขนาดใหญ่ที่มีปริมาณมาก มีความหลากหลาย และมีความเร็วสูงในการสร้างและการจัดการข้อมูล ซึ่งไม่สามารถจัดการได้ด้วยเครื่องมือหรือเทคนิคฐานข้อมูลแบบดั้งเดิม ใน Big Data มักจะมีคุณสมบัติที่เรียกว่า "3V" ดังนี้:

1.Volume (ปริมาณ):

- ข้อมูลมีปริมาณมาก ซึ่งสามารถวัดได้ในระดับเทราไบต์ (terabytes), เพตาไบต์ (petabytes), หรือแม้กระทั่งเอกซะไบต์ (exabytes)
- ปริมาณข้อมูลนี้เกิดจากการสะสมของข้อมูลจากแหล่งต่างๆ เช่น โซเชียลมีเดีย, เซนเซอร์, การทำธุรกรรมออนไลน์ และการเก็บข้อมูลจากอุปกรณ์ IoT

2.Variety (ความหลากหลาย):

- ข้อมูลมีความหลากหลายในรูปแบบ เช่น ข้อมูลที่มีโครงสร้าง (structured data), ข้อมูลที่ไม่มีโครงสร้าง (unstructured data), และข้อมูลกึ่งโครงสร้าง (semi-structured data) ตัวอย่างเช่น ข้อความ, ภาพถ่าย, วิดีโอ, เสียง, และข้อมูลจากเซนเซอร์ต่างๆ

3. Velocity (ความเร็ว):

- ข้อมูลมีความเร็วในการสร้างและการประมวลผลที่สูง ซึ่งต้องการการประมวลผลแบบเรียลไทม์หรือใกล้เคียงกับเวลาจริง
- ตัวอย่างเช่น การสตรีมข้อมูลจากเซนเซอร์, การวิเคราะห์ข้อมูลโซเชียลมีเดียทันที, และการทำธุรกรรมทางการเงินออนไลน์

นอกจากนี้ยังมีคุณสมบัติอื่น ๆ ที่เกี่ยวข้องกับ Big Data เช่น:

4.Veracity (ความถูกต้อง):

- ความถูกต้องและความน่าเชื่อถือของข้อมูลเป็นสิ่งสำคัญ เนื่องจากข้อมูลที่มีปริมาณมากอาจมีความผิดพลาดหรือความไม่แน่นอน

5.Value (คุณค่า):

- ข้อมูลมีคุณค่าที่สามารถนำไปใช้ในการวิเคราะห์เพื่อสร้างมูลค่าให้กับองค์กร เช่น การเพิ่มประสิทธิภาพการทำงาน การทำความเข้าใจลูกค้า และการตัดสินใจที่ดีมากยิ่งขึ้น.

แหล่งอ้างอิง:

1.Marr, B. (2015). *Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*. Wiley.

2.Jacobs, A. (2009). *The Pathologies of Big Data*. Communications of the ACM, 52(8), 36-44.

3.Big Data: A Revolution That Will Transform How We Live, Work, and Think by Viktor Mayer-Schönberger and Kenneth Cukier (2013).