คำชื่นชม

โลกของข้อมูลได้วิวัฒน์มาช่วงหนึ่งแล้ว จากนักออกแบบ สู่ผู้ดูแลฐานข้อมูล สู่ CIO จนเป็นสถาปนิกข้อมูล และหนังสือเล่มนี้ได้ส่งสัญญาณถึงขั้นต่อไปของวิวัฒนาการ จึงเป็นหนังสือที่ผู้อยู่ในสายงานวิศวกรรมข้อมูลจำเป็นต้องอ่าน

—Bill Inmon, ผู้ให้กำเนิดระบบ Data Warehouse

หนังสือเล่มนี้ให้คำแนะนำที่ยอดเยี่ยมสำหรับองค์กรธุรกิจ ในการเคลื่อนย้าย, ประมวลผล และจัดการกับข้อมูล เหมาะอย่างยิ่งสำหรับผู้สนใจงานวิศวกรรมข้อมูลและการวิเคราะห์ข้อมูล

—Jordan Tigani, ผู้ก่อตั้งและ CEO ของ MotherDuck
และวิศวกรผู้ก่อตั้งและผู้ร่วมพัฒนา BigQuery

หากต้องการเป็นผู้นำในอุตสาหกรรมของคุณ วิศวกรข้อมูลคือศูนย์กลางของการเปลี่ยนแปลง ซึ่งหนังสือเล่มนี้จะช่วยไขข้อข้องใจเกี่ยวกับวิศวกรรมข้อมูล และจะเป็นแนวทางสำคัญที่ทำให้คุณประสบความสำเร็จกับข้อมูล

—Bruno Aziza, หัวหน้าฝ่ายวิเคราะห์ข้อมูล Google Cloud

นี่เป็นหนังสือที่ยอดเยี่ยมมาก! Joe และ Matt จะให้คำตอบต่อคำถามที่ว่า “ฉันต้องเข้าใจอะไรบ้าง จึงจะทำงานด้านวิศวกรรมข้อมูลได้?”

—Andy Petrella, ผู้ก่อตั้ง Kensu

นี่คือหนังสือด้านวิศวกรรมข้อมูลที่ขาดหายไป เป็นรายละเอียดที่ครอบคลุมสิ่งที่จำเป็นสำหรับการเป็นวิศวกรข้อมูลที่ดี ฉันขอแนะนำให้ผู้ศึกษาด้านข้อมูลต้องอ่านผลงานของ Joe และ Matt

—Sarah Krasnik, หัวหน้าฝ่ายวิศวกรรมข้อมูล

หนังสือเล่มนี้ให้ภาพรวมพื้นฐานที่ยอดเยี่ยมเกี่ยวกับสถาปัตยกรรม, แนวทาง, วิธีการ และรูปแบบต่างๆ ที่ผู้ทำงานกับข้อมูลทุกคนต้องทราบ ในเล่มเต็มไปด้วยความรู้, คำแนะนำถึงแนวทางปฏิบัติ และสิ่งที่ต้องพิจารณาเมื่อต้องตัดสินใจเกี่ยวกับวิศวกรรมข้อมูล

—Veronika Durgin, หัวหน้าฝ่ายข้อมูลและการวิเคราะห์

ฉันรู้สึกเป็นเกียรติที่ Joe และ Matt ขอให้ช่วยตรวจสอบทางเทคนิคกับผลงานชิ้นเอกนี้ พวกเขาอธิบายส่วนประกอบสำคัญๆ ได้อย่างยอดเยี่ยม ด้วยสไตล์การเขียนทำให้เข้าใจได้ง่ายและครบถ้วน

—Chris Tabb, ผู้ร่วมก่อตั้ง LEIT DATA

นี่คือหนังสือเล่มแรกที่เจาะลึก และครอบคลุมความต้องการของวิศวกรข้อมูลในปัจจุบัน Joe และ Matt แสดงให้เห็นถึงความเชี่ยวชาญด้านวิศวกรรมข้อมูล และทำให้ผู้อ่านเข้าถึงได้ ไม่ว่าคุณจะเป็นผู้จัดการ, วิศวกรข้อมูลที่มีประสบการณ์ หรือผู้ที่ต้องการเข้าสู่วงการนี้ หนังสือเล่มนี้จะให้ข้อมูลเชิงปฏิบัติการเกี่ยวกับภูมิทัศน์วิศวกรรมข้อมูลในปัจจุบัน

—Jon King, หัวหน้าทีมสถาปนิกข้อมูล

มีสองสิ่งที่จะยังเกี่ยวข้องกับวิศวกรข้อมูลในปี 2042 ได้แก่ SQL และหนังสือเล่มนี้ ไม่ว่าคุณกำลังเริ่มต้น หรือกำลังยกระดับความรู้ Fundamentals of Data Engineering จะช่วยวางรากฐานสำหรับการเป็นผู้เชี่ยวชาญให้กับคุณ

—Kevin Hu, CEO, Metaplane

หนังสือเล่มนี้อัดแน่นไปด้วยข้อมูลที่ช่วยให้คุณเข้าใจถึงข้อดีข้อเสีย และตัดสินใจได้ดีที่สุดเมื่อออกแบบสถาปัตยกรรมข้อมูล และนำโซลูชันไปใช้ตลอดวงจรชีวิตวิศวกรรมข้อมูล

—Julie Price, ผู้จัดการผลิตภัณฑ์อาวุโส, SingleStore

หนังสือเล่มนี้เป็นทั้งบทเรียนประวัติศาสตร์, ทฤษฎี และความรู้จากประสบการณ์หลายสิบปีของ Joe และ Matt จึงสมควรอยู่ในชั้นวางหนังสือของมืออาชีพด้านข้อมูลทุกคน

—Scott Breitenother, ผู้ก่อตั้งและซีอีโอ, Brooklyn Data Co.

ไม่มีหนังสือเล่มใดที่จะครอบคลุมถึงการเป็นวิศวกรข้อมูลได้เท่านี้ Joe และ Matt เจาะลึกถึงความรับผิดชอบ, ผลกระทบ, การเลือกสถาปัตยกรรม และอื่นๆ อีกมากมาย ซึ่งแม้ว่าจะพูดถึงหัวข้อที่ซับซ้อนเหล่านี้ แต่เนื้อหาก็อ่านง่ายและเข้าใจง่าย เป็นการผสมผสานที่ทรงพลังจริงๆ

—Danny Leybzon, MLOps Architect

อยากให้มีหนังสือเล่มนี้ตั้งแต่ผมเริ่มทำงานกับกลุ่มวิศวกรข้อมูล เนื้อหาที่ครอบคลุมด้านต่างๆ อย่างกว้างขวาง ทำให้เห็นและเข้าใจบทบาทที่เกี่ยวข้องกับการสร้างวินัยด้านข้อมูลอย่างชัดเจน

—Tod Hansmann, รองประธานฝ่ายวิศวกรรม

หนังสือคลาสสิกที่คนในสาขาวิศวกรรมข้อมูลต้องอ่าน เพื่อเติมเต็มช่องว่างในฐานความรู้ปัจจุบัน คุณจะได้รับความเข้าใจในแนวคิดพื้นฐาน และข้อมูลเชิงลึกเกี่ยวกับบริบททางวิศวกรรมข้อมูล

—Matthew Sharp, วิศวกรข้อมูล และ ML

วิศวกรรมข้อมูลเป็นรากฐานของการวิเคราะห์, โมเดล ML และผลิตภัณฑ์ข้อมูลทุกประเภท แต่มีแหล่งข้อมูลน้อยมาก (ถ้ามี) ที่ให้มุมมองแบบองค์รวมเกี่ยวกับการเป็นวิศวกรข้อมูล หนังสือเล่มนี้จะวางรากฐานให้วิศวกรข้อมูลประสบความสำเร็จ และมีประสิทธิภาพ นี่คือหนังสือที่ผมขอแนะนำให้กับทุกคนที่ต้องการทำงานกับข้อมูล

—Tobias Macey, ผู้ดำเนินการ The Data Engineering Podcast

คำนำ

ที่มาของหนังสือเล่มนี้มีประวัติย้อนหลังกลับไปตั้งแต่ผู้เขียนเริ่มการเดินทาง จากงานด้านวิทยาศาสตร์ข้อมูลมาสู่งานด้านวิศวกรรมข้อมูล ซึ่งเรามักเรียกตัวเองแบบติดตลกว่า “นักวิทยาศาสตร์ข้อมูลที่ฟื้นคืนชีพ” เราทั้งคู่เคยมีประสบการณ์ในการทำโครงการด้านวิทยาศาสตร์ข้อมูลหลายโครงการ ซึ่งต้องดิ้นรนเพื่อให้โครงการเหล่านี้สำเร็จ เนื่องจากขาดรากฐานที่เหมาะสม แล้วการเดินทางของพวกเราสู่วิศวกรรมข้อมูลก็เริ่มต้นขึ้น เมื่อได้รับหน้าที่ด้านวิศวกรรมข้อมูลเพื่อสร้างรากฐานและโครงสร้างพื้นฐานของระบบข้อมูล

เมื่อวิทยาศาสตร์ข้อมูลได้รับความนิยม บริษัทต่างๆ ก็ทุ่มเงินมหาศาลให้กับบุคลากรด้านนี้ เพื่อหวังว่าจะได้รับผลตอบแทนที่คุ้มค่า แต่บ่อยครั้งที่นักวิทยาศาสตร์ข้อมูลต้องพบกับปัญหาพื้นฐานที่ตัวเองไม่เคยมีประสบการณ์ เช่น การรวบรวมและคลีนข้อมูล, การแปลง และโครงสร้างพื้นฐานของข้อมูล ซึ่งปัญหาเหล่านี้คือหน้าที่ด้าน “วิศวกรรมข้อมูล”

หนังสือเล่มนี้ไม่ได้พูดถึงอะไร

ก่อนที่จะพูดถึงว่า หนังสือเล่มนี้พูดถึงอะไร ผู้เขียนขอพูดถึงสิ่งที่หนังสือเล่มนี้ไม่ได้กล่าวถึงก่อนดีกว่า หนังสือเล่มนี้ไม่ได้พูดถึงวิศวกรรมข้อมูลโดยการใช้เครื่องมือ, เทคโนโลยี หรือแพลตฟอร์มที่เฉพาะเจาะจง (ซึ่งจะทำให้เป็นหนังสือที่มีอายุใช้งานสั้น) เราจึงมุ่งเน้นไปยังแนวคิดพื้นฐานที่อยู่เบื้องหลังงานวิศวกรรมข้อมูล

หนังสือเล่มนี้มีเนื้อหาเกี่ยวกับอะไร

แนวคิดหลักของหนังสือเล่มนี้คือ วงจรชีวิตของวิศวกรรมข้อมูล ซึ่งประกอบด้วย การสร้างข้อมูล, การจัดเก็บ, การนำข้อมูลเข้าสู่ระบบ, การแปลง และการให้บริการ ตั้งแต่ยุคเริ่มต้นของข้อมูล เราได้เห็นการเปลี่ยนแปลงของเทคโนโลยีและผลิตภัณฑ์มากมาย อย่างไรก็ตามขั้นตอนของวงจรชีวิตวิศวกรรมข้อมูลยังคงไม่เปลี่ยนแปลงไปมากนัก ดังนั้นการอ้างอิงจากวงจรชีวิตจะทำให้ผู้อ่านเข้าใจการนำเทคโนโลยีไปใช้กับปัญหาทางธุรกิจในโลกแห่งความเป็นจริงได้อย่างถ่องแท้ โดยผู้เขียนมีเป้าหมายที่จะสรุปให้เห็นภาพหลักการของงานวิศวกรรมข้อมูล ที่ยังสามารถยึดเป็นแนวทางได้อีกอย่างน้อยหนึ่งทศวรรษในอนาคต ถึงแม้เทคโนโลยีและผลิตภัณฑ์จะเปลี่ยนไปตามกาลเวลาก็ตาม

สิ่งหนึ่งที่ควรทราบก็คือ เราให้ความสำคัญกับระบบคลาวด์เป็นอันดับแรก เพราะเป็นการพัฒนาที่สร้างความเปลี่ยนแปลงไปอย่างสิ้นเชิง และจะคงอยู่ต่อไปอีกหลายทศวรรษ โดยระบบข้อมูลภายในองค์กรส่วนใหญ่จะย้ายไปที่คลาวด์ในที่สุด อย่างไรก็ตามแนวคิดส่วนใหญ่ในหนังสือเล่มนี้ก็สามารถใช้กับสภาพแวดล้อมที่ไม่ใช่ระบบคลาวด์ได้ด้วย

ใครควรอ่านหนังสือเล่มนี้

กลุ่มเป้าหมายหลักของหนังสือเล่มนี้ ได้แก่ ผู้ทำงานด้านเทคนิค, วิศวกรซอฟต์แวร์, นักวิทยาศาสตร์ข้อมูล และนักวิเคราะห์ ที่สนใจจะก้าวเข้าสู่สายงานวิศวกรรมข้อมูล หรือวิศวกรข้อมูลที่ทำงานเฉพาะด้าน แต่ต้องการพัฒนามุมมองที่ครอบคลุมมากขึ้น

สำหรับกลุ่มเป้าหมายรองของเรา ได้แก่ ผู้มีส่วนเกี่ยวข้องด้านข้อมูล ซึ่งทำงานร่วมกับผู้ปฏิบัติงานด้านเทคนิค เช่น หัวหน้าทีมวิศวกรข้อมูลที่มีพื้นฐานทางเทคนิค หรือผู้อำนวยการฝ่ายคลังข้อมูลที่ต้องการย้ายระบบไปยังคลาวด์

สิ่งที่ควรรู้ก่อน

ผู้อ่านควรมีความคุ้นเคยกับระบบข้อมูลประเภทต่างๆ ที่พบในองค์กรเป็นอย่างดี มีความคุ้นเคยกับ SQL และ Python (หรือภาษาโปรแกรมอื่นๆ) และมีประสบการณ์กับบริการคลาวด์

บริการบนคลาวด์ให้โอกาสในการได้รับประสบการณ์จริงกับเครื่องมือด้านข้อมูล ผู้เขียนขอแนะนำให้วิศวกรข้อมูลมีบัญชีของบริการคลาวด์ เช่น AWS, Azure, Google Cloud Platform, Snowflake, Databricks ซึ่งมีตัวเลือกให้ลองใช้งานฟรี แต่ผู้อ่านต้องระมัดระวังเรื่องค่าใช้จ่ายที่จะเกิดขึ้น และทดลองทำงานกับข้อมูลเพียงเล็กน้อย หรือใช้คลัสเตอร์แบบโหนดเดียวเพื่อศึกษาการทำงาน

สิ่งที่ได้จากหนังสือเล่มนี้

เนื้อหาในหนังสือเล่มนี้ มีเป้าหมายเพื่อช่วยสร้างรากฐานที่มั่นคงสำหรับการจัดการข้อมูล เมื่ออ่านจบ คุณจะเข้าใจถึง

- การจัดการข้อมูลจะส่งผลต่อนักวิทยาศาสตร์ข้อมูล, วิศวกรซอฟต์แวร์ หรือหัวหน้าทีมข้อมูล อย่างไร
- วิธีเลือกเทคโนโลยี, สถาปัตยกรรมข้อมูล และกระบวนการที่เหมาะสม
- การใช้วงจรชีวิตการจัดการข้อมูล เพื่อออกแบบและสร้างสถาปัตยกรรมที่แข็งแกร่ง
- แนวทางปฏิบัติสำหรับแต่ละขั้นตอนในวงจรชีวิตข้อมูล

และผู้อ่านจะสามารถ

- นำหลักการจัดการข้อมูลมาใช้กับหน้าที่ของนักวิทยาศาสตร์ข้อมูล, นักวิเคราะห์, วิศวกรซอฟต์แวร์, หัวหน้าทีมข้อมูล ฯลฯ
- เชื่อมโยงเทคโนโลยีคลาวด์ต่างๆ เข้าด้วยกัน เพื่อตอบสนองความต้องการของผู้ใช้ข้อมูลปลายทาง
- ประเมินปัญหาด้านการจัดการข้อมูล ด้วยกรอบแนวทางปฏิบัติแบบครบวงจร
- รวมการกำกับดูแลข้อมูล และการรักษาความปลอดภัย เข้ากับวงจรชีวิตการจัดการข้อมูล

สารบัญ

ส่วนที่ 1 รากฐานและบล็อกโครงสร้าง

บทที่ 1 Data Engineering คืออะไร? 25
ความหมายและคำจำกัดความ 25
วงจรชีวิตของวิศวกรรมข้อมูล 27
ประวัติความเป็นมาของวิศวกรรมข้อมูล 29
วิศวกรรมข้อมูลและวิทยาศาสตร์ข้อมูล 36
ทักษะและงานของ Data Engineering 38
วิศวกรข้อมูลกับความสมบูรณ์ของข้อมูล 39
พื้นฐานการศึกษาและทักษะของวิศวกรข้อมูล 44
มุมมองทางธุรกิจ 45
มุมมองทางเทคนิค 46
ความต่อเนื่องของบทบาท Data Engineering จาก A สู่ B 50
Data Engineer ภายในองค์กร 51
การเผชิญกับปัจจัยภายในและภายนอกองค์กร 51
วิศวกรข้อมูลกับบทบาททางเทคนิคด้านอื่นๆ 53
วิศวกรข้อมูลและความเป็นผู้นำทางธุรกิจ 57
สรุป 61
แหล่งข้อมูลเพิ่มเติม 61

บทที่ 2 วงจรชีวิตของวิศวกรรมข้อมูล 65
วงจรชีวิตของวิศวกรรมข้อมูลคืออะไร? 65
วงจรชีวิตข้อมูลต่างจากวงจรชีวิตวิศวกรรมข้อมูลอย่างไร 67
การเกิดขึ้นของข้อมูล: ระบบต้นทาง 67
พื้นที่จัดเก็บ (Storage) 70
การนำเข้าข้อมูล 72
การเปลี่ยนรูปหรือการแปลงข้อมูล 75
การให้บริการข้อมูล 76

งานแฝงในวงจรชีวิตวิศวกรรมข้อมูล 79
ความปลอดภัยของข้อมูล 80
การจัดการข้อมูล 81
DataOps 87
สถาปัตยกรรมข้อมูล 90
การประสานและกำหนดทิศทางข้อมูล 90
วิศวกรรมซอฟต์แวร์ 92
สรุป 94
แหล่งข้อมูลเพิ่มเติม 95

บทที่ 3 การออกแบบสถาปัตยกรรมข้อมูลที่ดี 99
สถาปัตยกรรมข้อมูลคืออะไร 99
นิยามของสถาปัตยกรรมองค์กร 99
คำนิยามของสถาปัตยกรรมข้อมูล 101
สถาปัตยกรรมข้อมูล “ที่ดี” 103
หลักการของสถาปัตยกรรมข้อมูลที่ดี 103
หลักข้อที่ 1: เลือกองค์ประกอบทั่วไปอย่างชาญฉลาด 105
หลักข้อที่ 2: วางแผนรับมือกับความผิดพลาด 105
หลักข้อที่ 3: สถาปัตยกรรมสำหรับการปรับขนาด 106
หลักข้อที่ 4: ใช้สถาปัตยกรรมเป็นเครื่องนำทาง 107
หลักข้อที่ 5: มีวิวัฒนาการอยู่เสมอ 107
หลักข้อที่ 6: สร้างระบบที่เชื่อมโยงกันแบบหลวมๆ 107
หลักข้อที่ 7: การตัดสินใจแบบย้อนกลับได้ 109
หลักข้อที่ 8: ให้ความสำคัญกับความปลอดภัย 109
หลักข้อที่ 9: เปิดรับ FinOps 111
แนวคิดหลักเกี่ยวกับสถาปัตยกรรม 112
โดเมนและบริการ 112
ระบบแบบกระจาย, ความสามารถในการปรับขนาด และการออกแบบ
เพื่อรับมือกับความผิดพลาด 113
เชื่อมต่อกันแบบแน่นหนา และแบบหลวม 115
การเข้าถึงของผู้ใช้: ผู้ใช้เดียว และหลายผู้ใช้ 119

สถาปัตยกรรมที่ขับเคลื่อนด้วยเหตุการณ์ 119
โปรเจ็กต์แบบบราวน์ฟิลด์ และแบบกรีนฟิลด์ 120
ตัวอย่างและประเภทของสถาปัตยกรรมข้อมูล 121
คลังข้อมูล 121
Data Lake 124
Data Lake รุ่นต่อไป และ Data Platform 125
สแต็กข้อมูลยุคใหม่ 125
สถาปัตยกรรมแลมบ์ดา 126
สถาปัตยกรรมแคปปา 126
โมเดล Dataflow และการรวมแบตช์กับสตรีม 127
สถาปัตยกรรมสำหรับ IoT 127
Data Mesh 130
ตัวอย่างสถาปัตยกรรมข้อมูลแบบอื่นๆ 131
ใครต้องเกี่ยวข้องกับการออกแบบสถาปัตยกรรมข้อมูลบ้าง? 132
สรุป 132
แหล่งข้อมูลเพิ่มเติม 132

บทที่ 4 การเลือกเทคโนโลยีในวงจรชีวิตวิศวกรรมข้อมูล 139
ขนาดและความสามารถของทีม 140
ความเร็วในการเข้าสู่ตลาด 140
ความสามารถในการทำงานร่วมกัน 141
ต้นทุนที่เหมาะสม และมูลค่าทางธุรกิจ 141
ต้นทุนรวมในการเป็นเจ้าของ 141
ค่าเสียโอกาสรวมในการเป็นเจ้าของ 142
FinOps 142
เทคโนโลยีคงตัว หรือเทคโนโลยีชั่วคราว 143
คำแนะนำจากผู้เขียน 144
ที่ตั้ง 144
ติดตั้งภายในองค์กร 145
คลาวด์ 145
ไฮบริดคลาวด์ 149

มัลติคลาวด์ 149
การกระจายศูนย์: Blockchain และ Edge 150
คำแนะนำจากผู้เขียน 150
ข้อโต้แย้งเรื่องการย้ายงานจากคลาวด์กลับระบบเดิม 151
พัฒนาขึ้นเอง หรือซื้อมาใช้ 152
ซอฟต์แวร์โอเพ่นซอร์ส 153
โซลูชันจากผู้ขาย 156
คำแนะนำจากผู้เขียน 157
ระบบหลักที่รวมเป็นชิ้นเดียว หรือระบบที่ประกอบจากส่วนย่อยๆ 157
ระบบที่รวมเป็นชิ้นเดียว 158
ระบบแบบโมดูลาร์ 159
ระบบเดียวแบบกระจาย 160
คำแนะนำจากผู้เขียน 160
มีเซิร์ฟเวอร์ หรือไม่มีเซิร์ฟเวอร์ 161
ไร้เซิร์ฟเวอร์ 161
คอนเทนเนอร์ 162
วิธีประเมินการมีเซิร์ฟเวอร์ เทียบกับแบบไร้เซิร์ฟเวอร์ 162
คำแนะนำจากผู้เขียน 164
การปรับแต่ง, ประสิทธิภาพ และเกณฑ์วัด 165
Big Data...สำหรับทศวรรษที่ 1990 165
การเปรียบเทียบต้นทุนแบบไร้ประโยชน์ 166
การปรับแต่งแบบไม่สมดุล 166
ข้อพึงระวังของผู้ซื้อ 166
งานแฝงและผลกระทบต่อการเลือกเทคโนโลยี 166
การจัดการข้อมูล 167
DataOps 167
สถาปัตยกรรมข้อมูล 167
ตัวอย่าง Orchestration: Airflow 167
วิศวกรรมซอฟต์แวร์ 168
สรุป 168
แหล่งข้อมูลเพิ่มเติม 168

ส่วนที่ 2 เจาะลึกวงจรชีวิตวิศวกรรมข้อมูล

บทที่ 5 การสร้างข้อมูลที่ระบบต้นทาง 173

แหล่งข้อมูล: ข้อมูลถูกสร้างขึ้นอย่างไร? 174

ระบบต้นทาง: แนวคิดหลัก 174

ไฟล์และข้อมูลแบบไม่มีโครงสร้าง 174

API 175

ฐานข้อมูลของแอพพลิเคชัน 175

ระบบประมวลผลเชิงวิเคราะห์ออนไลน์ (OLAP) 177

Change Data Capture (CDC) 177

Log 177

Log ของฐานข้อมูล 178

CRUD 179

การเพิ่มเข้าไปเท่านั้น 179

ข้อความและสตรีม 180

ชนิดของเวลา 181

รายละเอียดการปฏิบัติงานของระบบต้นทาง 182

ฐานข้อมูล 182

API 189

การใช้ข้อมูลร่วมกัน 190

แหล่งข้อมูลจากภายนอก 191

Message Queue และ Event-Streaming Platform 191

คุณต้องร่วมงานกับใครบ้าง 195

งานแฝงและผลกระทบต่อระบบต้นทาง 196

ความปลอดภัย 197

การจัดการข้อมูล 197

DataOps 198

สถาปัตยกรรมข้อมูล 199

การประสานและกำหนดทิศทางข้อมูล 200

วิศวกรรมซอฟต์แวร์ 200

สรุป 201

แหล่งข้อมูลเพิ่มเติม 202

บทที่ 6 พื้นที่จัดเก็บข้อมูล 205
องค์ประกอบพื้นฐาน 207
ดิสก์ไดรว์แม่เหล็ก 207
โซลิดสเตทไดรว์ 209
หน่วยความจำ RAM 210
เน็ตเวิร์ค และ CPU 210
Serialization 211
การบีบอัด 211
การแคช 212
ระบบจัดเก็บข้อมูล 212
จัดเก็บบนเครื่องเดียว หรือเก็บแบบกระจาย 213
ความสอดคล้องแบบ “ท้ายที่สุด” หรือแบบ “เข้มงวด” 213
พื้นที่จัดเก็บไฟล์ 214
บล็อกสตอเรจ 216
พื้นที่เก็บข้อมูลออบเจ็กต์ 219
ระบบจัดเก็บข้อมูลบนแคชและหน่วยความจำ 223
Hadoop Distributed File System (HDFS) 224
พื้นที่เก็บข้อมูลแบบสตรีมมิ่ง 225
อินเด็กซ์, การแบ่งพาร์ทิชัน และการทำคลัสเตอร์ 225
สตอเรจเชิงนามธรรมในวิศวกรรมข้อมูล 227
คลังข้อมูล 228
Data Lake 228
Data Lakehouse 228
แพลตฟอร์มข้อมูล 229
สถาปัตยกรรมที่จัดเก็บข้อมูลแบบสตรีมไปเป็นแบตช์ 229
แนวคิดและเทรนด์ของสตอเรจ 230
แคตตาล็อกข้อมูล 230
การแชร์ข้อมูล 231
สคีมา 231

แยกการประมวลผลออกจากที่เก็บข้อมูล 232
วงจรชีวิตของที่จัดเก็บข้อมูลและระยะการเก็บรักษา 234
ที่เก็บข้อมูลสำหรับผู้เช่ารายเดียว และสำหรับผู้เช่าหลายราย 237
คุณต้องร่วมงานกับใครบ้าง 238
งานแฝงเกี่ยวกับพื้นที่จัดเก็บข้อมูล 238
ความปลอดภัย 238
การจัดการข้อมูล 238
DataOps 239
สถาปัตยกรรมข้อมูล 240
การประสานและกำหนดทิศทางข้อมูล 240
วิศวกรรมซอฟต์แวร์ 240
สรุป 240
แหล่งข้อมูลเพิ่มเติม 241

บทที่ 7 การนำเข้าข้อมูล 243
การนำเข้าข้อมูลคืออะไร? 244
ข้อควรพิจารณาสำหรับขั้นตอนการนำเข้าข้อมูล 245
ข้อมูลนี้ มีหรือไม่มีขอบเขต 245
ความถี่ 246
การนำเข้าแบบซิงโครนัส และแบบอะซิงโครนัส 247
Serialization และ Deserialization 248
ปริมาณงานที่ทำได้ต่อหน่วยเวลา และความสามารถในการปรับขนาด 248
ความน่าเชื่อถือและความคงทน 249
เพย์โหลด (Payload) 249
รูปแบบพุช, พูล และโพล 252
ข้อควรพิจารณาสำหรับการนำเข้าแบบแบตช์ 253
สแนปช็อตหรือการดึงส่วนต่าง 254
การส่งออกและนำเข้าด้วยไฟล์ 255
ETL กับ ELT 255
การแทรก, อัพเดต และขนาดของแบตช์ 255
การย้ายข้อมูล 256

ข้อควรพิจารณาในการนำเข้าข้อมูลแบบข้อความและสตรีม 256
วิวัฒนาการของสคีมา 257
ข้อมูลล่าช้า 257
การจัดลำดับ และการจัดส่งหลายครั้ง 257
รีเพลย์ 258
Time to Live 258
ขนาดของข้อความ 258
การจัดการข้อผิดพลาด และคิวของเหตุการณ์ที่มีปัญหา 259
การพุชและพูลของผู้รับข้อมูล 259
ที่ตั้ง 260
วิธีนำเข้าข้อมูล 260
เชื่อมต่อกับฐานข้อมูลโดยตรง 260
การเก็บข้อมูลที่มีการเปลี่ยนแปลง 261
API 262
คิวข้อความและแพลตฟอร์มการสตรีมอีเวนต์ 263
ตัวเชื่อมต่อข้อมูลที่จัดการได้ 264
การย้ายข้อมูลด้วยพื้นที่เก็บออบเจ็กต์ 264
EDI 265
ฐานข้อมูลและการเอ็กซ์พอร์ตไฟล์ 265
ปัญหาในทางปฏิบัติเกี่ยวกับรูปแบบไฟล์ทั่วไป 265
เชลล์ 266
SSH 266
SFTP และ SCP 267
เว็บฮุค 267
เว็บอินเทอร์เฟซ 268
การดึงข้อมูลจากเว็บ 268
อุปกรณ์สำหรับถ่ายโอนข้อมูล 268
การแบ่งปันข้อมูล 269
คุณต้องทำงานร่วมกับใคร 269
ผู้มีส่วนได้ส่วนเสียทางอัพสตรีม 269
ผู้มีส่วนได้ส่วนเสียทางดาวน์สตรีม 270

งานแฝง 270
ความปลอดภัย 271
การจัดการข้อมูล 271
DataOps 272
Orchestration 273
วิศวกรรมซอฟต์แวร์ 273
สรุป 273
แหล่งข้อมูลเพิ่มเติม 274

บทที่ 8 คิวรี, โมเดล และการเปลี่ยนรูปข้อมูล 277
คิวรี 278
คิวรีคืออะไร? 278
การทำงานของคิวรี 279
ตัวเพิ่มประสิทธิภาพคิวรี 280
การปรับปรุงประสิทธิภาพคิวรี 280
คิวรีกับข้อมูลสตรีม 283
การสร้างแบบจำลองข้อมูล 288
โมเดลข้อมูลคืออะไร? 288
โมเดลข้อมูลเชิงแนวคิด, ลอจิก และกายภาพ 289
นอร์มัลไลเซชัน 290
เทคนิคสำหรับสร้างโมเดลการวิเคราะห์ข้อมูลแบบแบตช์ 294
โมเดลข้อมูลสตรีมมิ่ง 304
การเปลี่ยนรูปข้อมูล 305
การแปลงข้อมูลแบตช์ 306
Materialized View, Federation และ Query Virtualization 313
การแปลงและการประมวลผลข้อมูลสตรีมมิ่ง 315
คุณต้องทำงานร่วมกับใคร 317
ผู้มีส่วนได้ส่วนเสียทางอัพสตรีม 317
ผู้มีส่วนได้ส่วนเสียทางดาวน์สตรีม 318
งานแฝง 318
ความปลอดภัย 318

การจัดการข้อมูล 319
DataOps 319
สถาปัตยกรรมข้อมูล 319
การประสานและกำหนดทิศทางข้อมูล 320
วิศวกรรมซอฟต์แวร์ 320
สรุป 320
แหล่งข้อมูลเพิ่มเติม 321

บทที่ 9 การให้บริการข้อมูลสำหรับการวิเคราะห์, แมชชีนเลิร์นนิ่ง และ ETL แบบย้อนกลับ 325
ข้อควรคำนึงในการให้บริการข้อมูล 326
ความเชื่อมั่น 326
ผู้ใช้และกรณีการใช้งาน 327
ผลิตภัณฑ์ข้อมูล 328
ผู้ใช้ทำด้วยตัวเองหรือไม่? 329
นิยามและลอจิกของข้อมูล 330
Data Mesh 330
การวิเคราะห์ 331
การวิเคราะห์เชิงธุรกิจ 331
การวิเคราะห์เชิงปฏิบัติการ 333
การวิเคราะห์แบบฝังตัว 335
แมชชีนเลิร์นนิ่ง 335
สิ่งที่วิศวกรข้อมูลควรรู้เกี่ยวกับ ML 336
วิธีให้บริการข้อมูลสำหรับการวิเคราะห์และ ML 338
การส่งไฟล์ 338
ฐานข้อมูล 338
ระบบสตรีมมิ่ง 339
คิวรีกลุ่ม 340
การใช้ข้อมูลร่วมกัน 341
เลเยอร์ความหมาย และเลเยอร์ตัวชี้วัด 341
การให้บริการข้อมูลในโน้ตบุ๊ค 342

ETL แบบย้อนกลับ 344
คุณต้องทำงานร่วมกับใคร 345
งานแฝง 346
ความปลอดภัย 346
การจัดการข้อมูล 347
DataOps 348
สถาปัตยกรรมข้อมูล 348
Orchestration 348
วิศวกรรมซอฟต์แวร์ 349
สรุป 350
แหล่งข้อมูลเพิ่มเติม 350

ส่วนที่ 3 การรักษาความปลอดภัย, ความเป็นส่วนตัว และอนาคตของวิศวกรรมข้อมูล

บทที่ 10 การรักษาความปลอดภัย และความเป็นส่วนตัว 355
บุคคล 356
พลังแห่งการคิดเชิงลบ 356
จงหวาดระแวงอยู่เสมอ 356
กระบวนการ 357
ความปลอดภัยปลอมๆ กับความปลอดภัยจนเป็นนิสัย 357
การรักษาความปลอดภัยเชิงรุก 357
หลักการให้สิทธิ์น้อยที่สุด 358
ความรับผิดชอบร่วมกันในระบบคลาวด์ 358
สำรองข้อมูลอยู่เสมอ 358
ตัวอย่างนโยบายด้านความปลอดภัย 359
เทคโนโลยี 361
แพตช์และการอัพเดต 361
การเข้ารหัส 361
การลงบันทึก, เฝ้าระวัง และแจ้งเตือน 362
การเข้าถึงเครือข่าย 363
การรักษาความปลอดภัยสำหรับวิศวกรรมข้อมูลระดับล่าง 364

สรุป 365
แหล่งข้อมูลเพิ่มเติม 365

บทที่ 11 อนาคตของวิศวกรรมข้อมูล 367
วงจรชีวิตวิศวกรรมข้อมูลจะยังคงอยู่ 367
เครื่องมือด้านข้อมูลที่ซับซ้อนน้อยลง และใช้งานง่ายขึ้น 368
ระบบปฏิบัติการข้อมูลบนคลาวด์ และการทำงานร่วมกันที่ดีขึ้น 369
งานวิศวกรรมข้อมูลระดับ “องค์กรขนาดใหญ่” 371
ตำแหน่งและความรับผิดชอบจะแปรเปลี่ยนไป 371
ก้าวจาก Modern Data Stack สู่ Live Data Stack 372
Live Data Stack 373
ไปป์ไลน์สตรีมมิ่งและฐานข้อมูลการวิเคราะห์แบบเรียลไทม์ 374
การหลอมรวมของข้อมูลกับแอพพลิเคชัน 375
การรวมกันของแอพพลิเคชันและ ML 376
สสารมืดในโลกของข้อมูล 376
สรุป 377

ภาคผนวก ก รายละเอียดทางเทคนิคของการบีบอัดข้อมูล และซีเรียลไลเซชัน 379

ภาคผนวก ข เน็ตเวิร์คบนคลาวด์ 387