

สรุปบทเรียน หลักสูตร ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคลากร

(Big Data Analytics)

ระบบเรียนออนไลน์ e-Learning ของสถาบัน TDGA

ผู้รับการฝึกอบรม : นาย วชิรวิชัย เลิศสิทธิพันธ์

ตำแหน่ง : นิติกรปฏิบัติการ

สังกัด : กลุ่มนิติกร สำนักงานเลขาธิการกรม

หน่วยงานที่จัดอบรม : สถาบันพัฒนาบุคลากรภาครัฐด้านดิจิทัล Thailand Digital Government Academy

คำอธิบายบทเรียน

เรียนรู้เกี่ยวกับความหมายของ Big Data และการนำไปใช้ให้เกิดประโยชน์ กระบวนการในการจัดเก็บ Big Data รวมถึงเรียนรู้เกี่ยวกับเครื่องมือที่ใช้ในการวิเคราะห์ Big Data และแนวทางในการวิเคราะห์ Big Data เพื่อการทำงานภาครัฐ

วัตถุประสงค์

- เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data)
- เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่
- เพื่อให้ผู้เรียนมีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ

หัวข้อในบทเรียน

- ความหมายของ Big data
- คุณลักษณะ องค์ประกอบและประเภทของ Big Data
- กระบวนการทำงานของเครื่องมือวิเคราะห์ข้อมูล Hadoop
- Big Data Analytics กับการบริหารภาครัฐ

การพัฒนาความรู้ ๑ เรื่อง หลักสูตรความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการ และบุคลากร (Big Data) ผ่านระบบพัฒนาทักษะด้านดิจิทัล ของสถาบันพัฒนาบุคลากรภาครัฐด้านดิจิทัล Thailand Digital Government Academy สรุปสาระสำคัญจากการอบรม ดังนี้

บทที่ ๑ ความหมาย

Big data คือ การนำข้อมูลที่มีขนาดใหญ่ ปริมาณมากๆ มาผ่านการประมวลผล การวิเคราะห์ และแสดงผลด้วยวิธีที่เหมาะสม มีทั้งแบบโครงสร้างปกติ และโครงสร้างข้อมูลที่ไม่มีรูปแบบ ซึ่งทั้งหมดเป็นข้อมูลที่ใช้ในเชิงธุรกิจ มักใช้กับงานที่ต้องวิเคราะห์ ข้อมูลที่มีความซับซ้อน และไม่สามารถประเมินขนาดข้อมูลได้ ซึ่งต้องอาศัยการวิเคราะห์ข้อมูล (Big data Analytics) กับการบริหารภาครัฐ โดยรับฟังความเห็น รวบรวม และปรึกษากับผู้มีส่วนได้ส่วนเสีย วางแผนการลงทุนในการจัดโครงสร้าง มีความเข้าใจและมีทักษะทางธุรกิจ และทักษะทางเทคนิค เพื่อเตรียมความพร้อมภายใต้การเปลี่ยนแปลงของเทคโนโลยีและเจ้าหน้าที่ภาครัฐ ควรจะต้องปรับ mindset เพื่อให้สอดคล้องกับนโยบายของรัฐบาลในการอำนวยความสะดวกให้แก่ประชาชน

๑.๑ ขนาดของข้อมูล

ในปัจจุบันขนาดข้อมูลที่ถือว่ามีขนาดใหญ่มาก เริ่มต้นที่ขนาดมากกว่า Terabyte ขึ้นไป (เท่ากับ ๑,๐๒๔ Gigabyte) ถัดขึ้นไปก็เป็น Petabyte (เท่ากับ ๑,๐๒๔ Terabyte) และ Exabyte (เท่ากับ ๑,๐๒๔ Petabyte) โดยการจัดเก็บข้อมูลทั้งที่เป็นแบบ Structured หรือ Unstructured ก็ได้

๑.๒ รูปแบบของข้อมูลของ big data ก็สามารเป็นไปได้หลากหลาย ได้แก่

Behavioral data ข้อมูลเชิงพฤติกรรมการใช้งานต่างๆเช่น server log, พฤติกรรมการคลิก ดูข้อมูล, ข้อมูลการใช้ ATM เป็นต้น

Image & sounds ภาพถ่าย วิดีโอ รูปจาก google street view ภาพถ่ายทางการแพทย์ ลายมือ ข้อมูลเสียงที่ถูกบันทึกไว้ เป็นต้น

Languages text message ข้อความที่ถูก tweet เนื้อหาต่างๆในเว็บไซต์ เป็นต้น

Records ข้อมูลทางการแพทย์ ข้อมูลผลสำรวจที่มีขนาดใหญ่ ข้อมูลทางภาษี เป็นต้น

Sensors ข้อมูลอุณหภูมิจาก accelerometer ข้อมูลทางภูมิศาสตร์ เป็นต้น

บทที่ ๒ คุณลักษณะ องค์ประกอบและประเภทของ Big Data

๒.๑ คุณลักษณะของ Big Data

Volume ข้อมูลมีขนาดใหญ่ ปริมาณข้อมูลมาก ซึ่งสามารถเป็นได้ทั้งแบบ Offline หรือ Online

Variety ข้อมูลมีความหลากหลาย สามารถเป็นได้ทั้งที่มีโครงสร้างและข้อมูลที่ไม่สามารถจับ Pattern ได้

Velocity ข้อมูลมีการเปลี่ยนแปลงตลอดเวลาอย่างรวดเร็ว มีการส่งผ่านข้อมูลอย่างต่อเนื่องในลักษณะ Streaming ทำให้การวิเคราะห์ข้อมูลแบบ Manual มีข้อจำกัด

Varecity ข้อมูลมีความไม่ชัดเจน (Untrusted Uncleaned)

๒.๒ องค์ประกอบของ big data

๒.๒.๑ แหล่งข้อมูล (Data Source) แหล่งข้อมูลของ Big Data นั้นอาจมาจากหลากหลายแหล่ง เช่น โปรแกรม ข้อมูลธุรกรรม (transaction) อุปกรณ์ต่างๆ หรืออุปกรณ์ IOT ทำให้มีความยากในการจัดการข้อมูลเนื่องจากแต่ละแหล่งมีโครงสร้างข้อมูลที่แตกต่างกัน

๒.๒.๒ การนำเข้าข้อมูล (Ingestion) ซึ่งการส่งผ่านข้อมูลเป็นขั้นตอนแรกของ Big Data Ecosystem ในการดึงข้อมูลดิบ โดยข้อมูลดิบนั้นจะมาจากทั้ง แหล่งข้อมูลภายใน (internal sources) ฐานข้อมูลเชิงสัมพันธ์ (relational databases) ฐานข้อมูลที่ไม่ใช่เชิงสัมพันธ์ (non-relational databases) และแหล่งอื่นๆอีกมากมาย ซึ่งการนำเข้าข้อมูลจะมีหลักๆ ๒ ประเภท คือ Batch เป็นการรวบรวมและส่งผ่านกลุ่มข้อมูล Big Data ขนาดใหญ่ไปด้วยกัน โดยการใช้งานจะต้องถูกกระตุ้นโดยเงื่อนไข เช่น ตั้งเวลาไว้ หรือเรียกใช้เฉพาะกิจ และอีกประเภท คือ Streaming เป็นการไหลของข้อมูลแบบต่อเนื่อง ซึ่งจำเป็นอย่างมากสำหรับการวิเคราะห์ข้อมูลแบบ Real Time โดยในช่วงของการส่งผ่านข้อมูลจากแหล่งข้อมูลไปยังแหล่งเก็บข้อมูลอาจจะใช้กระบวนการได้หลากหลาย เช่น ETL ซึ่งย่อมาจาก Extract transform and load หรืออธิบายง่ายๆ ก็คือ ดึงข้อมูล แปลงข้อมูล และนำข้อมูลเข้าสู่แหล่งเก็บข้อมูลปลายทางนั่นเอง

๒.๓ แหล่งเก็บข้อมูล (Storage) การเก็บข้อมูลขนาดใหญ่ หรือ Big Data ในปัจจุบันมักจะเก็บไว้ใน Data Warehouse หรือ Data Lake ซึ่งอาจจะนำมาพักไว้เพื่อให้พร้อมนำไปใช้งานต่อในการวิเคราะห์หรือเก็บข้อมูลในอดีตก็ได้ โดย Data Warehouse จะเป็นการเก็บข้อมูล Big Data แบบดั้งเดิมที่เก็บได้เฉพาะข้อมูลที่มีโครงสร้างชัดเจน หรือกิ่งโครงสร้างจากหลายแหล่ง ส่วน Data Lake จะยืดหยุ่นกว่า และเก็บข้อมูลได้ทั้งข้อมูลที่มีโครงสร้างชัดเจนและข้อมูลที่ไม่มีโครงสร้างแน่นอน ซึ่งการเลือกใช้ก็ขึ้นอยู่กับเงื่อนไขหรือความต้องการในการนำ Big Data ไปใช้งานขององค์กร

๒.๔ การวิเคราะห์ (Analysis) เมื่อมีข้อมูลดิบในแหล่งข้อมูลแล้ว เราจะสามารถนำข้อมูล Big Data มาจัดเตรียมและทำความสะอาด (Data Cleansing) เพื่อเข้าสู่กระบวนการวิเคราะห์ข้อมูลได้ โดยกระบวนการวิเคราะห์ข้อมูลมักจะได้แก่

Descriptive Analytics หรือ สถิติเชิงพรรณนา เช่น ค่าเฉลี่ย ผลรวม

Diagnostics Analytics หรือ ความสัมพันธ์ของตัวแปร เช่น Correlation

Predictive Analytics หรือ การพยากรณ์หรือทำนายอนาคต เช่น Machine Learning

Prescriptive analytics หรือ การวิเคราะห์แบบให้คำแนะนำ เช่น Optimization ซึ่งการวิเคราะห์จะขึ้นอยู่กับชนิดและบริบทของข้อมูล Big Data นั้น รวมถึงสิ่งที่องค์กรอยากรู้เกี่ยวกับข้อมูล

๒.๕ การนำผลลัพธ์ไปใช้ (Consumption) ถือเป็นองค์ประกอบสำคัญของ Big Data ซึ่งการนำข้อมูล Big Data จากการวิเคราะห์ไปใช้อาจอยู่ในรูปแบบ Data Visualization เช่น ตาราง กราฟ Dashboard สวยๆ หรืออาจจะเป็น insight ที่ตอบคำถามทางธุรกิจ หรือส่งผลต่อการตัดสินใจขององค์กรได้

บทที่ ๓ Hadoop

Hadoop หรือชื่อเรียกอย่างเป็นทางการคือ Apache Hadoop เป็น software framework สำหรับจัดการกับชุดข้อมูลขนาดใหญ่ (large data set) และการประมวลผลแบบกระจาย (distributed processing) บนเครื่องคอมพิวเตอร์ที่เชื่อมต่อกันเป็นกลุ่ม (cluster)

๓.๑ หน้าทีของ Hadoop

๓.๑.๑ บริหารจัดการหน่วยเก็บข้อมูล (storage management)

๓.๑.๒ บริหารจัดการทรัพยากรสำหรับการประมวลผล (computing resource management)

บทที่ ๔ Big Data Analytics กับการบริหารภาครัฐ

จากแผนพัฒนาดิจิทัลเพื่อเศรษฐกิจและสังคมเพื่อปฏิรูปประเทศไทยสู่ DIGITAL THAILAND นั้น คือการพัฒนาประเทศให้ที่สามารถสร้างสรรค์ และใช้ประโยชน์จากเทคโนโลยีดิจิทัลอย่างเต็มศักยภาพ โดยการปรับปรุงและยกระดับโครงสร้างพื้นฐาน การสร้างสรรค์นวัตกรรม ระบบข้อมูล ทูมมนุษย์ และทรัพยากรอื่น เพื่อขับเคลื่อนเศรษฐกิจและสังคมของประเทศไปสู่ความมั่นคง มั่งคั่ง และยั่งยืนโดยแผนพัฒนาดิจิทัลเพื่อเศรษฐกิจและสังคม มีเป้าหมาย ๔ ประการดังนี้

๔.๑ เพิ่มขีดความสามารถในการแข่งขัน ก้าวทันเวทีโลก ด้วยการใช้นวัตกรรมและเทคโนโลยีดิจิทัลเป็นเครื่องมือหลักในการสร้างสรรค์นวัตกรรมการผลิตการบริการ

๔.๒ สร้างโอกาสทางสังคมอย่างเท่าเทียมด้วยข้อมูลข่าวสารและบริการต่างๆ ผ่านสื่อดิจิทัล เพื่อยกระดับคุณภาพชีวิตของประชาชน

๔.๓ พัฒนาทุนมนุษย์สู่ยุคดิจิทัล ด้วยการเตรียมความพร้อมให้บุคลากรทุกกลุ่มมีความรู้ และทักษะที่เหมาะสมต่อการดำเนินชีวิตและการประกอบอาชีพในยุคดิจิทัล

๔.๔ ปฏิรูปกระบวนการศึการทำงานและการให้บริการของภาครัฐด้วยเทคโนโลยีดิจิทัล และการใช้ประโยชน์จากข้อมูลเพื่อให้เกิดการปฏิบัติงานโปร่งใส มีประสิทธิภาพ และประสิทธิผล

จึงเรียนมาเพื่อโปรดพิจารณา หากเห็นชอบ ขอได้โปรดเสนอแจ้งกลุ่มสารบรรณ ดำเนินการในส่วนที่เกี่ยวข้องต่อไป