ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคลากรภาครัฐทุกระดับ

โดย นางสาวปุณยนุช ลิมเมธาพงศา
นักวิชาการแผนที่ภาพถ่ายปฏิบัติการ

วัตถุประสงค์

๑. เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data)
๒. เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่
๓. เพื่อให้ผู้เรียนมีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ

เนื้อหา

Big Data คือ ข้อมูลขนาดใหญ่ มีทั้งแบบโครงสร้างปกติและโครงสร้างข้อมูลที่ไม่มีรูปแบบ ซึ่งทั้งหมดเป็นข้อมูลที่ใช้ในเชิงธุรกิจ มักจะถูกใช้กับงานพวกที่ต้องวิเคราะห์ข้อมูลที่มีความซับซ้อนและไม่สามารถประเมินขนาดข้อมูลได้

รูปแบบของข้อมูล Big Data สามารถเป็นไปได้หลากหลาย โดยมี 5 แบบ ได้แก่

Behavioral Data
ได้แก่ ข้อมูลเชิงพฤติกรรมการใช้งานต่างๆ เช่น Server Log พฤติกรรมการคลิกดูข้อมูลหรือข้อมูลการใช้ ATM

Image & Sounds
ตัวอย่างเช่น ภาพถ่าย วีดีโอ รูปจาก Google Street View ภาพถ่ายทางการแพทย์ ลายมือ ข้อมูลเสียงที่ถูกบันทึกไว้

Languages
ตัวอย่างเช่น Text Message ข้อความที่ถูก Tweet เนื้อหาต่างๆในเว็บไซต์

Records
ตัวอย่างเช่น ข้อมูลทางการแพทย์ ข้อมูลผลสำรวจ-ที่มีขนาดใหญ่ ข้อมูลทางภาษี

Sensors
ตัวอย่างเช่น ข้อมูลอุณหภูมิ, Accelerometer, ข้อมูลทางภูมิศาสตร์

Big Data ประกอบด้วยคุณลักษณะ ๔ ประการ คือ

- Volume ข้อมูลมีขนาดใหญ่ มีปริมาณข้อมูลมาก ซึ่งสามารถเป็นได้ทั้งข้อมูลแบบ Offline หรือ Online
- Variety ข้อมูลมีความหลากหลาย สามารถเป็นได้ทั้งที่มีโครงสร้างและข้อมูลที่ไม่สามารถจับ Pattern ได้
- Velocity ข้อมูลมีการเปลี่ยนแปลงตลอดเวลาอย่างรวดเร็ว มีการส่งผ่านข้อมูลอย่างต่อเนื่องในลักษณะ Streaming ทำให้การวิเคราะห์ข้อมูลแบบ Manual มีข้อจำกัด
- Veracity ข้อมูลมีความไม่ชัดเจน (Untrusted, Uncleaned)

Data Lake คือ ที่เก็บข้อมูลขนาดใหญ่ที่สามารถเก็บข้อมูลได้ทุกรูปแบบจากหลายแหล่ง โดยข้อมูลที่จัดเก็บเป็นข้อมูลดิบจำนวนมากและมีขนาดใหญ่ ข้อมูลไม่มีรูปแบบที่แน่นอน การเข้าถึงข้อมูลไม่สามารถเข้าถึงได้ง่ายต้องใช้เจ้าหน้าที่วิเคราะห์ข้อมูล (Data Scientist) ที่มีความรู้ความสามารถเข้ามาจัดการข้อมูล

Data Lake เกิดขึ้นจากการนำเอาข้อมูลจากแหล่งข้อมูลภายนอกองค์กรมาใช้มากขึ้น ปริมาณข้อมูลจากแหล่งภายนอกมีการเพิ่มขึ้นอย่างต่อเนื่อง และเพื่อแก้ไขข้อจำกัดหลายอย่างของ Data warehouse ที่ใช้กันมานาน

Big Data Analytics การวิเคราะห์ข้อมูล Big Data ทำให้มีข้อมูลที่เป็นข้อเท็จจริงซึ่งผ่านการวิเคราะห์อย่างเป็นระบบเพื่อใช้ประกอบการตัดสินใจ โดยระดับการวิเคราะห์ก็เป็นได้หลากหลายแล้วแต่รูปแบบการนำไปใช้งาน แบ่งเป็น ๓ ระดับ ได้แก่

Descriptive Analytics

เป็นการวิเคราะห์ในระดับที่บอกว่าเกิดอะไรขึ้น จำนวนเท่าไร ถี่แค่ไหน เกิดเหตุการณ์สำคัญ-ตอนไหน ตรงไหนบ้าง

Predictive Analytics

เป็นการวิเคราะห์ในลักษณะที่ซับซ้อนขึ้นไปอีกขั้นหนึ่ง คือ เป็นการประเมินว่าจะเกิดอะไรขึ้นต่อไป มีการให้ข้อมูลตัวชี้วัดของผลลัพธ์ที่อาจจะเกิดขึ้น ถ้าแนวโน้มยังเป็นอย่างนี้ต่อไป

Prescriptive Analytics

เป็นรูปแบบการวิเคราะห์ในลักษณะที่ซับซ้อนและยากที่สุด เพราะไม่เพียงพยากรณ์หรือทำนายว่าจะเกิดอะไรขึ้น แต่ยังให้คำแนะนำในทางเลือกต่างๆ และผลของทางเลือกต่างๆ ว่าจะมีผลดีและผลเสียอย่างไร โมเดลของ Prescriptive Analytics นั้นจะสามารถปรับเปลี่ยนได้ตามข้อมูลที่เพิ่มเติมเข้ามามากขึ้น และ Prescriptive Analytics นี้ยังเป็นการใช้ข้อมูลที่มากที่สุดและเกี่ยวพันกับเรื่อง Big Data เป็นอย่างมาก

Method รูปแบบการวิเคราะห์ สามารถแบ่งเป็น ๓ ประเภท ได้แก่

- **Data Mining** คือ วิเคราะห์ข้อมูลจากข้อมูลจำนวนมาก (Big Data) เพื่อหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยทำการจำแนกประเภท รูปแบบ เชื่อมโยงข้อมูลที่มีความสัมพันธ์กันและหาความน่าจะเป็นที่จะเกิดขึ้น เพื่อให้ได้องค์ความรู้ใหม่ที่สามารถนำไปใช้ประกอบการตัดสินใจในด้านต่างๆ เช่น ตลาดหลักทรัพย์ทางธุรกิจ ทางด้านการแพทย์ ยุทธศาสตร์ทหาร เป็นต้น
- **Text Mining** คือ เป็นเทคนิคเพื่อค้นหารูปแบบ (Pattern) จากข้อความจำนวนมหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่องและการรู้จำแบบ หรือในนิยามหนึ่ง การทำเหมืองข้อความ คือ กระบวนการที่กระทำกับข้อความ (โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง หลักคณิตศาสตร์ หลักการประมวลเอกสาร (Document Processing) หลักการประมวลข้อความ (Text Processing) และการประมวลผลภาษาธรรมชาติ (Natural Language Processing)
- **Machine Learning** เป็นศาสตร์ของการสร้างโมเดลคณิตศาสตร์ มุ่งเน้นที่จะสร้างองค์ความรู้จากข้อมูล โดยเริ่มจากการสร้างโมเดลทางคณิตศาสตร์ที่มีความยืดหยุ่น และสามารถปรับตัวเองเข้ากับข้อมูลที่ได้รับ เพื่อที่จะสามารถทำนายอนาคตได้ เทคโนโลยีที่อยู่เบื้องหลังหุ่นยนต์อัจฉริยะต่างๆ หรือเทคโนโลยีการบินแบบไร้คนขับหรือระบบวิเคราะห์ภาพ เสียง ภาพเคลื่อนไหวและภาษามนุษย์

Social Media Command Center กำลังเป็นเครื่องมือสำคัญและกำลังได้รับความนิยมเป็นอย่างสูงที่จะคอยดูแลความเคลื่อนไหวธุรกิจองค์กรที่เกิดขึ้นบนโลกออนไลน์ การบริโภคสื่อออนไลน์จะทำให้ธุรกิจเห็นพฤติกรรมของผู้บริโภคที่เป็นกลุ่มเป้าหมาย รูปแบบที่ผู้บริโภคเข้าไปมีปฏิสัมพันธ์กับธุรกิจในโลกออนไลน์ คือ Data ที่มีค่าของธุรกิจทั้งการคลิก การกดแชร์ การใช้เวลากับหน้าเว็บไซต์แต่ละแห่ง โดยข้อมูลที่รวบรวมมาจากออนไลน์ ได้แก่ ข้อมูลด้าน Demographic หรืออายุ เพศ การศึกษา หรืออาชีพ ข้อมูลด้านไลฟ์สไตล์และความสนใจ ซึ่งธุรกิจจะนำไปใช้ในการเลือกกลุ่มเป้าหมายในการโฆษณาได้แม่นยำมากขึ้น

ธุรกิจสามารถใช้ประโยชน์จาก Social Media Command Center ได้ ดังนี้

- **Trend Watching** ตรวจสอบกระแสในขณะนั้น เพื่อดูว่าคนกล่าวถึงเรื่องใดบ้าง
- **Brand Health** ตรวจสอบสุขภาพของธุรกิจ มีการกล่าวถึงแง่ใดบ้าง บวกหรือลบมากกว่ากัน
- **Online Campaign Measurement** วัดกระแสของแคมเปญออนไลน์ของธุรกิจได้
- **Influencer Analysis** ตรวจสอบการพูดถึงธุรกิจ โดย Influencer
- **Event with Live Stream** ค้นหาจากแฮชแท็กของ Event แบบเรียลไทม์
- **Market Research** เพื่อให้ธุรกิจสามารถวิเคราะห์คู่แข่งและผู้บริโภคได้อย่างมีประสิทธิภาพ
- **Geomapping** ทำให้แบรนด์ตรวจสอบ Feedback จากผู้บริโภคที่อยู่ต่างสถานที่กันได้
- **Keyword Selection** เพื่อดูว่าผู้บริโภคเชื่อมโยงชื่อธุรกิจกับคำว่าอะไรบ้าง

Big Data Analytics กับการบริหารภาครัฐ องค์กรภาครัฐในยุคดิจิทัลมีความจำเป็นที่จะต้องสร้างมูลค่าในการวิเคราะห์ Big Data โดยมีแนวทาง ดังนี้

๑. รับฟังความเห็น รวบรวมข้อมูล และปรึกษากับผู้มีส่วนได้ส่วนเสีย

๒. วางแผนการลงทุนในการจัดโครงสร้าง

๓. มีความเข้าใจและมีทักษะทางธุรกิจและทักษะทางเทคนิค

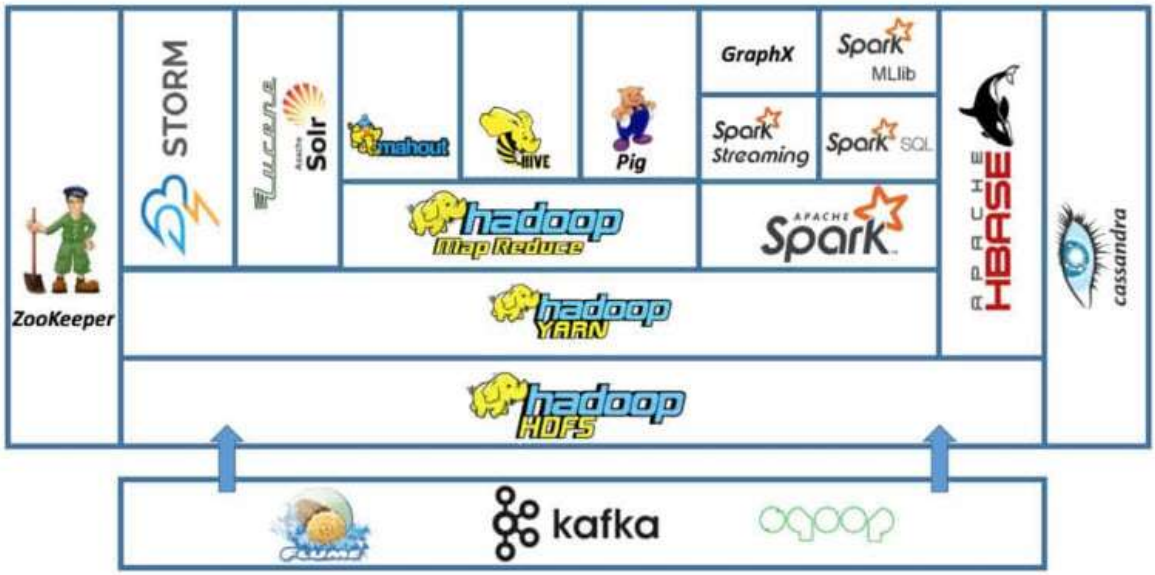
๔. เตรียมพร้อมภายใต้การเปลี่ยนแปลงของเทคโนโลยี

๕. เจ้าหน้าที่ภาครัฐจะต้องปรับ Mindset ในการเข้าร่วมกับทุกภาคส่วน

๖. ปรับปรุงวิธีคิดและกระบวนการเพื่อทำให้เกิดการแลกเปลี่ยนข้อมูลและการใช้ข้อมูลร่วมกันระหว่างหน่วยงานภาครัฐ

๗. กำหนดแนวทางและการบริการให้คำปรึกษาในด้าน Big Data Analytics ให้แก่ทุกภาคส่วน

Hadoop มีกระบวนการทำงานและเครื่องมือ ดังนี้



- Hive เป็นเครื่องมือสำหรับผู้ต้องการสืบค้น (Query) ข้อมูลที่เก็บใน HDFS ด้วยภาษาลักษณะ SQL แทนที่จะต้องมาเขียนโปรแกรม Map/Reduce โดย Hive จะทำหน้าที่ในการแปล SQL like ให้มาเป็น Map/Reduce แล้วก็ทำการรันแบบ Batch
- Pig เป็นเครื่องมือคล้ายๆกับ Hive ที่ช่วยให้ประมวลผลข้อมูลโดยไม่ต้องเขียนโปรแกรม Map/Reduce ซึ่ง Pig จะใช้โปรแกรมภาษา script ง่ายๆที่เรียกว่า Pig Latin แทน โดย Pigเหมาะกับการทำ ETL สำหรับการแปลงข้อมูลในรูปแบบต่างๆเช่น JSON
- Sqoop เป็นเครื่องมือในการถ่ายโอนข้อมูลระหว่างฐานข้อมูลที่อยู่รูปแบบ Table บน RDBMS อย่าง SQL server, Oracle หรือ MySQL กับข้อมูลบน HDFS ของ Hadoop

- Flume เป็นเครื่องมือในการดึงข้อมูลจากระบบอื่นๆแบบ Realtime เข้าสู่ HDFS เช่นการดึง Log จาก Web Server การดึงข้อมูลเหล่านี้จะต้องมีการติดตั้ง Agent ที่เครื่อง Server
- HBase เป็นเครื่องมือที่จะทำให้ Hadoop สามารถอ่านและเขียนข้อมูลแบบ Realtime Random Access ได้โดยจะทำให้เป็น BigTable ที่เก็บข้อมูลได้ไม่จำกัด row หรือ column ซึ่ง HBase ก็จะเป็นเสมือนการทำให้ Hadoop เป็น NoSQL Database
- Oozie เป็นเครื่องมือในการทำ Workflow จะช่วยให้เราเอาคำสั่งประมวลผลต่างๆของระบบ Hadoop เช่น Map/Reduce, Hive หรือ Pig มาเชื่อมต่อกันในรูปของ Workflow ได้
- Hue ย่อมาจากคำว่า Hadoop User Experience เป็นเครื่องมือช่วยทำ User interface ของ Hadoop ให้ใช้งานได้ง่ายขึ้นกว่าการต้องใช้ command line
- Mahout เป็นเครื่องมือของ Data Scientist ที่ต้องการทำ Predictive Analytics ข้อมูลบน Hadoop โดยใช้ภาษาจาวา ทั้งนี้ Mahout สามารถใช้ Algorithm ที่เป็น Recommender, Classification และ Clustering ได้

ประโยชน์ที่ได้รับ

ผู้เรียนเกิดความรู้ความเข้าใจพื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data) เข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ และได้รับความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่ โดยบทเรียนจากหลักสูตรความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคลากรภาครัฐทุกระดับนี้ สามารถนำไปประยุกต์ใช้ในการปฏิบัติงานของตนเองและหน่วยงาน