**สรุปรายงานการพัฒนาข้าราชการการฝึกอบรมการพัฒนาทางไกลด้วยระบบการฝึกอบรม**
**ผ่านสื่ออิเล็กทรอนิกส์ สถาบันพัฒนาบุคลาการด้านดิจิทัลภาครัฐ (TDGA e-Learning)**
**หลักสูตร ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูล สำหรับข้าราชการและบุคลากรภาครัฐทุกระดับ (Big Data)**

**โดย นางสาวทิวา ปาตีคำ นักวิชาการเกษตรชำนาญการพิเศษ กลุ่มวิชาการฯ**

## หลักสูตร ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูล สำหรับข้าราชการและบุคลากรภาครัฐทุกระดับ (Big Data)

โดย รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด
สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช

**วัตถุประสงค์**

1. เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data)
2. เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่
3. เพื่อให้ผู้เรียนมีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ

**ข้อมูลใหญ่ (Big Data)**

คือ ข้อมูลดิจิดัลที่จะถูกสร้างจากทุกที่และตลอดเวลา โดยข้อมูลข้อมูลขนาดใหญ่ มีทั้งแบบโครงสร้างปกติและโครงสร้างข้อมูลที่ไม่มีรูปแบบ ซึ่งทั้งหมดเป็นDATA ข้อมูลที่ใช้ในเชิงธุรกิจ มักจะถูกใช้กับงานพวกที่ต้องวิเคราะห์ ข้อมูลที่มีความซับซ้อน และไม่สามารถ ประเมินขนาดข้อมูลได้

**รูปแบบของข้อมูลของ Big data สามารถเป็นไปได้หลากหลาย ได้แก่**

1. Behavioral Data ได้แก่ ข้อมูลเซ็งพฤติกรรมการใช้งานต่างๆ เช่น Server Log พฤติกรรมการคลิกดู ข้อมูลหรือ ข้อมูลการใช้ ATM

2. Image & Sounds ตัวอย่างเช่น ภาพถ่าย วิดีไอ รูปจาก Google Street View ภาพถ่าย ทางการแพทย์ ลายมือ ข้อมูลเสียงที่ถูกบันทึกไว้

3. Languages ตัวอย่างเช่น Text Message ข้อความที่ถูก Tweet เนื้อหาต่างๆ ในเว็บไซต์ Record ตัวอย่างเช่น ข้อมูลทางการแพทย์ ข้อมูลผลสำรวจที่มีขนาดใหญ่ ข้อมูลทางภาษี

4. Sensors ตัวอย่างเช่น ข้อมูลอุณหภูมิ, Accelerometer, ข้อมูลทางภูมิศาสตร์

**Big Data ประกอบด้วยคุณลักษณะ 4 ประการ ได้แก่**

1. ปริมาณ (Volume) หมายถึง ข้อมูลมีขนาดใหญ่ มีปริมาณข้อมูลมากซึ่งสามารถเป็นได้ทั้งข้อมูลแบบ Offline หรือ Online ซึ่งโครงสร้างข้อมูลของระบบฐานข้อมูลไม่สามารถจัดเก็บข้อมูลได้

2. ความหลากหลาย (Variety) หมายถึง ข้อมูลมีความหลากหลาย สามารถเป็นได้ ทั้งที่มีโครงสร้างและข้อมูลที่ไม่สามารถจับ Pattern ได้ ได้แก่ รูปภาพ วีดีโอเพลง ข้อมูลจากระบบฐานข้อมูล หรือข้อมูลจากสื่อสังคมออนไลน์

3. ความเร็ว (Velocity) หมายถึง ข้อมูลมีการเปลี่ยนแปลงตลอดเวลาอย่างรวดเร็ว มีการส่งผ่านข้อมูลอย่างต่อเนื่องในลักษณะ Streaming ทำให้การวิเคราะห์ข้อมูลแบบ Manual มีข้อจำกัด

4. ความจริง (Veracity) หมายถึง ข้อมูลที่เป็นความจริง แม่นยำ ถึงข้อมูลจะมาจากแหล่งข้อมูลคนละที่ หรือคนละชนิด จึงจะต้องมีการจัดระเบียบและวิเคราะห์ว่าข้อมูลใดมีความถูกต้องแม่นยำมากที่สุด

**Data Lake** เกิดขึ้นเนื่องจาก การนำเอาข้อมูลจากแหล่งข้อมูลภายนอกองค์กร ข้อมูลจากเครือข่ายข้อมูลที่กระจายไปทั่วโลกมาใช้มากขึ้น ปริมาณข้อมูลจากแหล่งภายนอกเพิ่มอย่างต่อเนื่องและมีแนวโน้มที่จะเติบโตแบบก้าวกระโดดมากขึ้น และการแก้ไขข้อจำกัดหลายอย่างของ Data Warehouse ที่ใช้กันมานาน

ข้อมูลที่จัดเก็บ คือ

- ข้อมูลดิบจำนวนมากและมีขนาดใหญ่
- ข้อมูลไม่มีรูปแบบที่แน่นอน
- การเข้าถึงข้อมูลไม่สามารถเข้าถึงได้ง่ายต้องใช้ความสามารถของเจ้าหน้าที่วิเคราะห์ข้อมูล (Data Scientist)

ความแตกต่างระหว่าง Data Lake เมื่อเทียบกับ Data Warehouse

- เก็บข้อมูลทั้งหมด
- สนับสนุนข้อมูลทุกชนิด ไม่ใช่เพียงข้อมูลแบบ Structure
- ผู้ใช้ทุกประเภทสามารถใช้งานได้ ประมวลและวิเคราะห์ข้อมูลได้รวดเร็วกว่า

**Big Data Analytics**

การวิเคราะห์ข้อมูล Big Data ทำให้มีข้อมูลที่เป็นข้อเท็จจริงซึ่งผ่านการวิเคราะห์อย่างเป็นระบบเพื่อใช้ประกอบการตัดสินใจ โดยระดับของการวิเคราะห์ก็เป็นได้หลากหลายแล้วแต่รูปแบบการนำไปใช้งาน โดยแบ่งเป็น 3 ประเภท

1. การวิเคราะห์แบบพยากรณ์ (Predictive Analytics) เป็นการวิเคราะห์เพื่อพยากรณ์สิ่งที่กำลังจะเกิดขึ้นหรือน่าจะเกิดขึ้น โดยใช้ข้อมูลที่ได้เกิดขึ้นแล้วกับแบบจำลองทางสถิติ หรือ เทคโนโลยีปัญญาประดิษฐ์ต่างๆ (Artificial intelligence) ตัวอย่างเช่น การพยากรณ์ยอดขาย การพยากรณ์ผลประชามติ เป็นต้น

2. การวิเคราะห์ข้อมูลแบบพื้นฐาน (Descriptive Analytics) การวิเคราะห์ข้อมูลแบบพื้นฐาน เพื่อแสดงผลที่เกิดขึ้น หรือกำลังจะเกิดขึ้น จากข้อมูลในอดีต ในลักษณะที่เข้าใจง่ายสามารถสร้างขึ้นได้ด้วยตนเอง เช่น รายงาน แผนภูมิ กราฟ ตาราง เป็นต้น ซึ่งจะช่วยให้เข้าใจการเปลี่ยนแปลงที่เกิดขึ้นกับองค์กรได้ดียิ่งขึ้น

3. การวิเคราะห์แบบให้คำแนะนำ (Prescriptive Analytics) ป็นการวิเคราะห์ข้อมูลที่มีความซับซ้อนที่สุด เป็นทั้งการพยากรณ์สิ่งต่างๆ ที่จะเกิดขึ้น ข้อดี ข้อเสีย สาเหตุ และระยะเวลาของสิ่งที่จะเกิดขึ้น ร่วมถึงการให้คำแนะนำทางเลือกต่างๆ ที่มีอยู่ และผลของแต่ละทางเลือก

**Data Driven Business**

ปัจจุบันมีการนำข้อมูล ไปวิเคราะห์เพื่อนำไปใช้ในการตัดสินใจการทำธุรกิจอย่างแพร่หลาย สินทรัพย์ในทางธุรกิจ ข้อมูลถือเป็นทรัพย์สินที่มีมูลค่า ข้อมูลจะแสดงให้เห็นจุดอ่อนและจุดแข็งจากการดำเนินงานที่ผ่านมา และช่วยให้เกิดการพัฒนาธุรกิจ ได้แก่

- การเข้าถึงลูกค้าได้ดีขึ้น การเก็บข้อมูลของลูกค้า เพื่อให้ได้มาซึ่งบริการและสินค้าที่ตรงต่อความต้องการของลูกค้า ได้แก่ ต้องการช่องทางซื้อ-ขาย และช่องทางการชำระเงินที่สะดวก นอกจากนี้ยังทำให้แต่ละองค์กรเกิดความเปลี่ยนแปลงในด้านของการจัดการ โดยเฉพาะข้อมูลที่ต้องการมีการพัฒนาระบบจัดเก็บและระบบป้องกันรักษาข้อมูล นอกจากนี้ทางองค์กรยังต้องตื่นตัวในการสร้างและปรับปรุงนโยบายที่เกี่ยวกับการใช้ระบบข้อมูลให้มีความรัดกุมและทันสมัยอยู่เสมอ

- การพัฒนาประสิทธิภาพและการทำงาน Big Data มีความสามารถที่จะช่วยด้านการพัฒนาประสิทธิภาพและการทำงานภายในของธุรกิจเกือบทุกประเภท เช่น ติดตามการทำงานของพนักงาน การติดตามด้านสุขภาพและความเครียด ที่อาจเกิดขึ้นในระหว่างการทำงาน สามารถช่วยพัฒนาในการทรัพยากรบุคคลและการจ้างงานได้ด้วยนอกจากนี้ทางองค์กรยังต้องตื่นตัวในการสร้างและปรับปรุงนโยบายที่เกี่ยวกับการใช้ระบบข้อมูลให้มีความรัดกุมและทันสมัยอยู่เสมอ

- การพัฒนาความพึงพอใจของลูกค้าและผลิตภัณฑ์ องค์กรสามารถใช้ข้อมูลที่เก็บมาได้จากลูกค้าในการพัฒนาผลิตภัณฑ์และประสบการณ์การใช้สินค้าได้ เช่น การติดเซ็นเซอร์ที่จะช่วยให้องค์กรเข้าในถึงวิธีการใช้งานของลูกค้า เพื่อคาดการณ์ถึงปัญหาที่อาจจะเกิดขึ้นกับในอนาคต

**การบริโภคสื่อออนไลน์ (Social Media Command)**

จะทำให้ธุรกิจเห็นพฤติกรรมของผู้บริโภคที่เป็นกลุ่มเป้าหมาย รูปแบบที่ผู้บริโภคเข้าไปมีปฏิสัมพันธ์กับธุรกิจในโลกออนไลน์คือ Data ที่มีค่าของธุรกิจ ทั้งการคลิก การกดแชร์ การใช้เวลากับหน้าเว็บไซต์แต่ละแห่ง

ข้อมูลที่รวบรวมมาจากออนไลน์ ได้แก่ ข้อมูลด้าน Demographic หรืออายุ เพศ การศึกษา หรืออาชีพ และข้อมูลด้านไลฟ์สไตล์ และความสนใจ ซึ่งองค์กรจะนำไปใช้ในการเลือกกลุ่มเป้าหมายในการโฆษณาได้แม่นยำมากขึ้น

Social Media Command Center กำลังเป็นเครื่องมือสำคัญและกำลังได้รับความนิยมเป็นอย่างสูง ที่จะคอยดูแลความเคลื่อนไหวธุรกิจองค์กรที่เกิดขึ้นบนโลกออนไลน์ แบ่งเป็น

1. Data Visualization เป็น การแสดงผลข้อมูลในรูปแบบที่ เข้าใจได้ง่าย
2. Real Time Monitoring ไม่ใช่การรวบรวมเป็นรายงานสรุป แต่แสดงผลออกมาแบบเรียลไทม์
3. Quality Data นอกจากจะบอกปริมาณแล้ว ยังบอกทิศทางและรายละเอายดว่าคนกล่าวถึงองค์กรในแง่มุมบวกหรือลบ

**Big Data Analytics กับการบริหารภาครัฐ**

องค์กรภาครัฐในยุคดิจิตอลจำเป็นต้อง สร้างมูลค่าจากการวิเคราะห์ Big data โดยมีแนวทางดังนี้

1. รับฟังความเห็น รวบรวมข้อมูล และปรึกษากับผู้มีส่วนได้ส่วนเสีย
2. วางแผนการลงทุนในการจัดโครงสร้าง
3. มีความเข้าใจและมีทักษะทางธุรกิจและทักษะทางเทคนิค
4. เตรียมพร้อมภายใต้การเปลี่ยนแปลงของเทคโนโลยี
5. เจ้าหน้าที่ภาครัฐจะต้องปรับ Mindset ในการเข้าร่วมกับทุกภาคส่วน
6. ปรับปรุงวิธีคิดและกระบวนการเพื่อทำให้เกิดการแลกเปลี่ยนข้อมูล และการใช้ข้อมูลร่วมกันระหว่างหน่วยงานภาครัฐ
7. กำหนดแนวทางและการบริการให้คำปรึกษาในด้าน Big Data Analytics ให้แก่ทุกภาคส่วน

**เครื่องมือวิเคราะห์ข้อมูล (Apache Hadoop)**

อาปาเชฮาดูป (Apache Hadoop) หมายถึง ซอฟต์แวร์เฟรมเวิร์ค (Framework) ที่ถูกออกแบบมาเพื่อรองรับการทำงานบนระบบคอมพิวเตอร์แบบกระจาย (Distributed Computing) และสนับสนุนการประมวล (Processing) ที่มีความเสถียรสูงและสามารถพิ่มขยายจำนวนเครื่องในระบบไต้อย่งมหาศาล จึงเหมาะกับการประมวลแบบขนาน (Parallel Processing) ที่มีความเสถียรสูง และสามาถเพิ่มขยายจำนวนเครื่องในระบบได้อย่างมหาศาล จึงเหมาะกับการประมวลผลข้อมูลใหญ่ (Big Data) โดยฮาดูปเป็นซอฟต์แวร์แบบโอเพนซอร์ซ (Open Source) ของมูลนิธิอาปาเชซอฟต์แวร์ (Apache Software Foundation) ที่เปิดโอกาสให้บุคคลอื่นนำเอาซอฟต์แวร์นั้นไปพัฒนาต่อได้ ซึ่ง ฮาดูป ประกอบตัวยกลุ่มของชุดคำสั่งต่างๆ (Libraries) เพื่อช่วยอำนวยความสะดวกแก่นักพัฒนาแอปพลิเคชันที่จะสร้งระบบหรือวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics) ได้อย่างมีประสิทธิภาพ