

ผู้รับการฝึกอบรม : นายมานิต แยมมาค

ตำแหน่ง : เจ้าพนักงานการเกษตรชำนาญงาน

สังกัด : ศูนย์พัฒนาโครงการหลวงภาคเหนือ ฝ่ายวิชาการเพื่อการพัฒนาที่ดิน สำนักงานพัฒนาที่ดินเขต ๖

สรุปสาระสำคัญ “ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคลากรภาครัฐทุกระดับ”
ดังนี้

วัตถุประสงค์

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

ภาครัฐ

เนื้อหา

Big Data คือ ข้อมูลขนาดใหญ่ มีทั้งแบบโครงสร้างปกติและโครงสร้างข้อมูลที่ไม่มีรูปแบบ ซึ่งทั้งหมดเป็นข้อมูลที่ใช้ในเชิงธุรกิจมักจะถูกใช้กับงานพวกที่ต้องวิเคราะห์ข้อมูลที่มีความซับซ้อนและไม่สามารถประเมิน ขนาดข้อมูลได้

Big Data ประกอบด้วยคุณลักษณะ ๔ ประการ คือ

1. Volume ข้อมูลมีขนาดใหญ่ มีปริมาณข้อมูลมาก ซึ่งสามารถเป็นได้ทั้งข้อมูลแบบ Offline หรือ Online
2. Variety ข้อมูลมีความหลากหลาย สามารถเป็นได้ทั้งที่มีโครงสร้างและข้อมูลที่ไม่สามารถจับ Pattern ได้
3. Velocity ข้อมูลมีการเปลี่ยนแปลงตลอดเวลาอย่างรวดเร็ว มีการส่งผ่านข้อมูลอย่างต่อเนื่อง ในลักษณะ Streaming ทำให้การวิเคราะห์ข้อมูลแบบ Manual มีข้อจำกัด
๔. 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 โดยระดับการวิเคราะห์เป็นได้หลากหลายแล้วแต่รูปแบบการนำไปใช้งาน แบ่งเป็น ๓ ระดับ ได้แก่

๑. Descriptive Analytics เป็นการวิเคราะห์ในระดับที่บอกว่าเกิดอะไรขึ้น จำนวนเท่าไร ที่ไหน เกิดเหตุการณ์สำคัญตอนไหนตรงไหนบ้าง

๒. Predictive Analytics เป็นการวิเคราะห์ในลักษณะที่ซับซ้อนขึ้นไปอีกขั้นหนึ่ง คือเป็นการประเมินว่าจะเกิดอะไรขึ้น ต่อไป มีการให้ข้อมูลตัวชี้วัดของผลลัพธ์ที่อาจจะเกิดขึ้นถ้าแนวโน้มยังเป็นอย่างนี้ต่อไป

๓. Prescriptive Analytics เป็นรูปแบบการวิเคราะห์ในลักษณะที่ซับซ้อนและยากที่สุด เพราะไม่เพียงพยากรณ์หรือทำนายว่า จะเกิดอะไรขึ้น แต่ยังให้คำแนะนำในทางเลือกต่างๆ และผลของทางเลือกต่างๆว่า จะมีผลดีและผลเสียอย่างไร โมเดลของ Prescriptive Analytics นั้นจะสามารถปรับเปลี่ยนได้ตามข้อมูลที่เพิ่มเติมเข้ามามากขึ้น และ Prescriptive Analytics นี้ยังเป็นการใช้ข้อมูลที่มากที่สุดและเกี่ยวข้องกับเรื่อง Big Data เป็นอย่างมาก

Method รูปแบบการวิเคราะห์ สามารถแบ่งเป็น ๓ ประเภท ได้แก่

๑. Data Mining คือ วิเคราะห์ข้อมูลจากข้อมูลจำนวนมาก (Big Data) เพื่อหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยทำการจำแนกประเภทรูปแบบเชื่อมโยงข้อมูลที่มีความสัมพันธ์กันและหาความน่าจะเป็นที่จะเกิดขึ้น เพื่อให้ได้องค์ความรู้ใหม่ที่สามารถนำไปใช้ประกอบการตัดสินใจในด้านต่างๆ เช่น ตลาดหลักทรัพย์ทางธุรกิจ ทางด้านการแพทย์ ยุทธศาสตร์ทหาร เป็นต้น

๒. Text Mining คือ เป็นเทคนิคเพื่อค้นหารูปแบบ (Pattern) จากข้อความจำนวนมากโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติการเรียนรู้ของเครื่องและการรู้จำแบบหรือในนิยามหนึ่ง การทำเหมืองข้อความ คือกระบวนการที่กระทำกับข้อความ (โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง หลักคณิตศาสตร์ หลักการประมวลเอกสาร (Document Processing) หลักการประมวลข้อความ (Text Processing) และการประมวลผลภาษาธรรมชาติ (Natural Language Processing)

๓. Machine Learning เป็นศาสตร์ของการสร้างโมเดลคณิตศาสตร์ มุ่งเน้นที่จะสร้างองค์ความรู้จากข้อมูล โดยเริ่มจากการสร้างโมเดลทางคณิตศาสตร์ที่มีความยืดหยุ่น และสามารถปรับตัวเองเข้ากับข้อมูลที่ได้รับ เพื่อที่จะสามารถทำนายอนาคตได้เทคโนโลยีที่อยู่เบื้องหลังหุ่นยนต์อัจฉริยะต่างๆ หรือเทคโนโลยีการบินแบบไร้คนขับหรือระบบวิเคราะห์ภาพ เสียง ภาพเคลื่อนไหวและภาษามนุษย์

Social Media Command Center กำลังเป็นเครื่องมือสำคัญและกำลังได้รับความนิยมเป็น

อย่างสูงที่จะคอยดูแลความเคลื่อนไหวธุรกิจองค์กรที่เกิดขึ้นบนโลกออนไลน์ การบริโภคสื่อออนไลน์จะทำให้ธุรกิจเห็นพฤติกรรมของผู้บริโภคที่เป็นกลุ่มเป้าหมาย รูปแบบที่ผู้บริโภคเข้าไปมีปฏิสัมพันธ์กับธุรกิจในโลกออนไลน์ คือ Data ที่มีค่าของธุรกิจทั้งการคลิก การกดแชร์ การใช้เวลากับหน้าเว็บไซต์แต่ละแห่ง โดยข้อมูลที่รวบรวม มาจากออนไลน์ ได้แก่ ข้อมูลด้าน Demographic หรืออายุ เพศ การศึกษา หรืออาชีพ ข้อมูลด้านไลฟ์สไตล์ และความสนใจ ซึ่งธุรกิจจะนำไปใช้ในการเลือกกลุ่มเป้าหมายในการโฆษณาได้แม่นยำมากขึ้น

ธุรกิจสามารถใช้ประโยชน์จาก Social Media Command Center ได้ ดังนี้

1. Trend Watching ตรวจสอบกระแสในขณะนั้น เพื่อดูว่าคนกล่าวถึงเรื่องใดบ้าง
2. Brand Health ตรวจสอบสุขภาพของธุรกิจ มีการกล่าวถึงแง่ใดบ้าง บวกหรือลบมากกว่ากัน
3. Online Campaign Measurement วัดกระแสของแคมเปญออนไลน์ของธุรกิจได้
4. Influencer Analysis ตรวจสอบการพูดถึงธุรกิจ โดย Influencer
5. Event with Live Stream ค้นหาจากแฮชแท็กของ Event แบบเรียลไทม์
6. Market Research เพื่อให้ธุรกิจสามารถวิเคราะห์คู่แข่งและผู้บริโภคได้อย่างมีประสิทธิภาพ
7. Geomapping ทำให้แบรนด์ตรวจสอบ Feedback จากผู้บริโภคที่อยู่ต่างสถานที่กันได้
8. Keyword Selection เพื่อดูว่าผู้บริโภคเชื่อมโยงชื่อธุรกิจกับคำว่าอะไรบ้าง

Big Data Analytics กับการบริหารภาครัฐ องค์กรภาครัฐในยุคดิจิทัลมีความจำเป็นที่จะต้อง

สร้างมูลค่าในการวิเคราะห์ Big Data โดยมีแนวทาง ดังนี้

1. รับฟังความเห็น รวบรวมข้อมูล และปรึกษากับผู้มีส่วนได้ส่วนเสีย
2. วางแผนการลงทุนในการจัดโครงสร้าง
3. มีความเข้าใจและมีทักษะทางธุรกิจและทักษะทางเทคนิค
4. เตรียมพร้อมภายใต้การเปลี่ยนแปลงของเทคโนโลยี
5. เจ้าหน้าที่ภาครัฐจะต้องปรับ Mindset ในการเข้าร่วมกับทุกภาคส่วน
6. ปรับปรุงวิธีคิดและกระบวนการเพื่อให้เกิดการแลกเปลี่ยนข้อมูล และการใช้ข้อมูลร่วมกัน

ระหว่างหน่วยงานภาครัฐ

7. กำหนดแนวทางและการบริการให้คำปรึกษาในด้าน Big Data Analytics ให้แก่ทุกภาคส่วน

Hadoop มีกระบวนการทำงานและเครื่องมือ ดังนี้

1. Hive เป็นเครื่องมือสำหรับผู้ต้องการสืบค้น (Query) ข้อมูลที่เก็บใน HDFS ด้วยภาษาลักษณะ

SQL แทนที่จะต้องมาเขียนโปรแกรม Map/Reduce โดย Hive จะทำหน้าที่ในการแปล SQL like ให้มาเป็น Map/Reduce แล้วก็ทำการรันแบบ Batch

๒. Pig เป็นเครื่องมือคล้ายๆกับ Hive ที่ช่วยให้ประมวลผลข้อมูลโดยไม่ต้องเขียนโปรแกรม Map/Reduce ซึ่ง Pig จะใช้โปรแกรมภาษา script ง่ายๆที่เรียกว่า Pig Latin แทน โดย Pig เหมาะกับการทำ ETL สำหรับการแปลงข้อมูลในรูปแบบต่างๆเช่น JSON
๓. Sqoop เป็นเครื่องมือในการถ่ายโอนข้อมูลระหว่างฐานข้อมูลที่อยู่รูปแบบ Table บน RDBMS อย่าง SQL server, Oracle หรือ MySQL กับข้อมูลบน HDFS ของ Hadoop
๔. Flume เป็นเครื่องมือในการดึงข้อมูลจากระบบอื่นๆแบบ Realtime เข้าสู่ HDFS เช่นการดึง Log จาก Web Server การดึงข้อมูลเหล่านี้จะต้องมีการติดตั้ง Agent ที่เครื่อง Server
๕. HBase เป็นเครื่องมือที่จะทำให้ Hadoop สามารถอ่านและเขียนข้อมูลแบบ Realtime Random Access ได้โดยจะทำให้เป็น BigTable ที่เก็บข้อมูลได้ไม่จำกัด row หรือ column ซึ่ง HBase ก็จะเป็นเสมือนการทำให้ Hadoop เป็น NoSQL Database
๖. Oozie เป็นเครื่องมือในการทำ Workflow จะช่วยให้เราเอาคำสั่งประมวลผลต่างๆของระบบ Hadoop เช่น Map/Reduce, Hive หรือ Pig มาเชื่อมต่อกันในรูปของ Workflow ได้
๗. Hue ย่อมาจากคำว่า Hadoop User Experience เป็นเครื่องมือช่วยทำ User interface ของ Hadoop ให้ใช้งานได้ง่ายขึ้นกว่าการต้องใช้ command line
๘. 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 การวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคลากรภาครัฐทุกระดับนี้ สามารถนำไปประยุกต์ใช้ในการ ปฏิบัติงานของตนเองและหน่วยงาน