ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคคลากรภาครัฐทุกระดับ (BIG DATA)

โดยนายสราวุธ อัคพิน
นักวิชาการแผนที่ภาพถ่ายปฏิบัติการ

วัตถุประสงค์

๑. เพื่อให้มีความรู้พื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data)
๒. เพื่อให้มีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่
๓. เพื่อให้มีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ

เนื้อหา

ปัจจุบันการแข่งขันกันทางข้อมูลค่อนข้างสูง ใครมีข้อมูลมากกว่าย่อมได้เปรียบ การที่มีข้อมูลเป็นจำนวนมาก และหลากหลายอาจก่อให้เกิดปัญหาในการใช้งาน ดังนั้นจึงจำเป็นต้องมีการจัดการ จำแนก รวมถึงการวิเคราะห์ข้อมูล เพื่อให้องค์กรสามารถเลือก และนำข้อมูลไปใช้งานได้อย่างรวดเร็ว ที่สำคัญคือการนำเสนอข้อมูลที่เป็นจำนวนมากนั้นต้องเข้าใจได้ง่าย เพื่อให้เท่าทันต่อการแข่งขันในยุคของข้อมูล

Big Data คือ ข้อมูลขนาดใหญ่ มีทั้งแบบโครงสร้างปกติ และโครงสร้างข้อมูลที่ไม่มีรูปแบบ ซึ่งทั้งหมดเป็นข้อมูลที่ใช้ในเชิงธุรกิจ มักจะถูกใช้กับงานพวกที่ต้องวิเคราะห์ ข้อมูลที่มีความซับซ้อน และไม่สามารถประเมินขนาดข้อมูลได้ รูปแบบของข้อมูล Big Data ได้แก่

๑. Behavioral Data เช่น พฤติกรรมการคลิกดู
๒. Image & Sounds เช่น ภาพถ่าย วิดีโอ รูปจากGoogle Street View ข้อมูลเสียงที่ถูกบันทึกไว้
๓. Languages เช่น Text Message ข้อความที่ถูก Tweet เนื้อหาต่างๆ ในเว็บไซต์
๔. Records เช่น ข้อมูลทางการแพทย์ ข้อมูลผล สำรวจที่มีขนาดใหญ่ ข้อมูลทางภาษี
๕. Sensors เช่น ข้อมูลอุณหภูมิ ข้อมูลทางภูมิศาสตร์

Big Data ประกอบด้วยคุณลักษณะ ๔ ประการ คือ

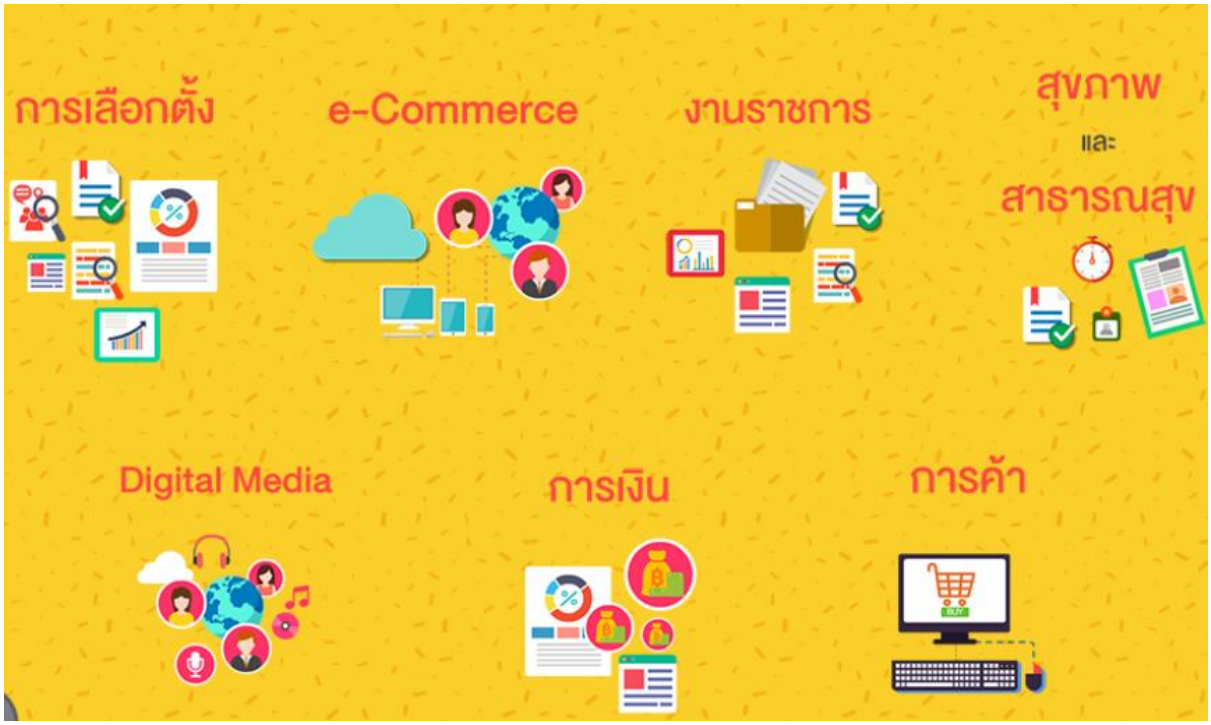
๑. Volume ข้อมูลมีขนาดใหญ่ มีปริมาณข้อมูลมาก ซึ่งสามารถเป็นได้ทั้งข้อมูลแบบ offline หรือ online
๒. Variety ข้อมูลมีความหลากหลาย เป็นได้ทั้งที่มีโครงสร้างและข้อมูลไม่สามารถจับ Pattern ได้
๓. Velocity ข้อมูลมีการเปลี่ยนแปลงตลอดเวลา มีการส่งผ่านข้อมูลอย่างต่อเนื่อง ทำให้การวิเคราะห์ข้อมูลแบบ Manual มีข้อจำกัด
๔. Veracity ข้อมูลมีความไม่ชัดเจน(Untrusted, Uncleaned)

การวิเคราะห์ข้อมูล Big data ทำให้ข้อมูลที่เป็นข้อเท็จจริงซึ่งผ่านการวิเคราะห์อย่างเป็นระบบเพื่อใช้ประกอบการตัดสินใจ โดยระดับการวิเคราะห์ก็เป็นได้หลากหลาย แล้วแต่รูปแบบการนำไปใช้งาน ดังนี้

Descriptive Analytics เป็นการวิเคราะห์ในระดับที่บอกว่าเกิดอะไรขึ้น จำนวนเท่าไหร่ ถี่แค่ไหน เกิดเหตุการณ์สำคัญตอนไหน ตรงไหนบ้าง

Predictive Analytics เป็นการวิเคราะห์ในลักษณะที่ซับซ้อนขึ้นไปอีกขั้น คือ เป็นการประเมินว่าจะเกิดอะไรขึ้นต่อไป มีการให้ข้อมูลตัวชี้วัดของผลลัพธ์ที่อาจจะเกิดขึ้นถ้าแนวโน้มยังเป็นอย่างนี้ต่อไป

Prescriptive Analytics เป็นรูปแบบการวิเคราะห์ข้อมูลที่มีความซับซ้อนและยากที่สุด เพราะไม่เพียงพยากรณ์หรือทำนายว่าอะไรจะเกิดขึ้น แต่ยังให้คำแนะนำในทางเลือกต่างๆ และผลของทางเลือกต่างๆ ว่าจะมีผลดีและผลเสียอย่างไร โมเดลของ Prescriptive Analytics นั้นจะสามารถปรับเปลี่ยนได้ตามข้อมูลที่เพิ่มเติมเข้ามามากขึ้น และ Prescriptive Analytics นี้ยังเป็นการใช้ข้อมูลที่มากที่สุด และเกี่ยวพันกับเรื่อง Big Data เป็นอย่างมาก



รูปภาพ แสดงตัวอย่างการใช้ Big Data

Big data Analytics กับการบริหารงานภาครัฐ ในยุคดิจิทัล องค์กรภาครัฐต้องสร้างมูลค่าจากการวิเคราะห์ข้อมูล Big data ตามแนวทาง ดังนี้

๑. รับฟังความเห็น รวบรวมข้อมูล และปรึกษากับผู้มีส่วนได้ส่วนเสีย

๒. วางแผนการลงทุนในการจัดโครงสร้าง

๓. มีความเข้าใจและมีทักษะทางธุรกิจและทักษะทางเทคนิค

๔. เตรียมพร้อมภายใต้การเปลี่ยนแปลงของเทคโนโลยี

๕. เจ้าหน้าที่ภาครัฐจะต้องปรับ Mindset ในการเข้าร่วมกับทุกภาคส่วน

๖. ปรับปรุงวิธีคิดและกระบวนการเพื่อทำให้เกิดการแลกเปลี่ยนข้อมูลและ การใช้ข้อมูลร่วมกันระหว่างหน่วยงานภาครัฐ

๗. กำหนดแนวทางและการบริการให้คำปรึกษาในด้าน Big DataAnalytics ให้แก่ทุกภาคส่วน

Big data Analytics เป็นเครื่องมือที่มีความจำเป็นของภาครัฐในทุกประเทศ ในการขับเคลื่อนการบริหารราชการแผ่นดินอย่างมีประสิทธิภาพและขับเคลื่อนเศรษฐกิจให้ทันต่อการเปลี่ยนแปลงทางเทคโนโลยีอย่างก้าวกระโดดในศตวรรษที่ ๒๑

รูปแบบในการวิเคราะห์ แบ่งเป็น ๓ รูปแบบ ดังนี้

Data Mining คือ เป็นการวิเคราะห์เพื่อหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยการจำแนกประเภทรูปแบบ เชื่อมโยงข้อมูลที่มีความสัมพันธ์กัน และหาความน่าจะเป็นที่จะเกิดขึ้น เพื่อให้ได้องค์ความรู้ใหม่นำไปใช้ประกอบการตัดสินใจในด้านต่างๆ เช่น ตลาดหลักทรัพย์ ทางด้านการแพทย์

Text Mining เป็นเทคนิคเพื่อค้นหารูปแบบ (Pattern) จากข้อความจำนวนมหาศาล โดยใช้ขั้นตอนวิธีจากวิชาสถิติ เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่

Machine Learning เป็นศาสตร์ของการสร้างโมเดลคณิตศาสตร์ มุ่งเน้นที่จะสร้างองค์ความรู้จากข้อมูล โดยเริ่มจากการสร้างโมเดลทางคณิตศาสตร์ที่มีความยืดหยุ่นและสามารถปรับตัวเองเข้ากับข้อมูลที่ได้รับ เพื่อจะสามารถทำนายอนาคตได้

การวิเคราะห์ข้อมูลด้วยเครื่องมือ Hadoop

เป็นซอฟต์แวร์ (Software) แบบโอเพนซอร์ช (Open Source) ที่เข้ามาช่วยเราจัดการเก็บและประมวลผลข้อมูล Big Data ได้ Hadoop ประกอบไปด้วยกลุ่มของชุดคำสั่งต่างๆ (Libraries) โดยข้อมูล (Data) ใส่เข้าไปใน Hadoop (เครื่องมือ) ทำการวิเคราะห์ เกิดเป็นผลลัพธ์เอาไปใช้งานเพื่อช่วยอำนวยความสะดวกในการวิเคราะห์ข้อมูลขนาดใหญ่ ได้อย่างมีประสิทธิภาพ

การทำงานของ Hadoop ประกอบด้วย ๔ ส่วนหลัก

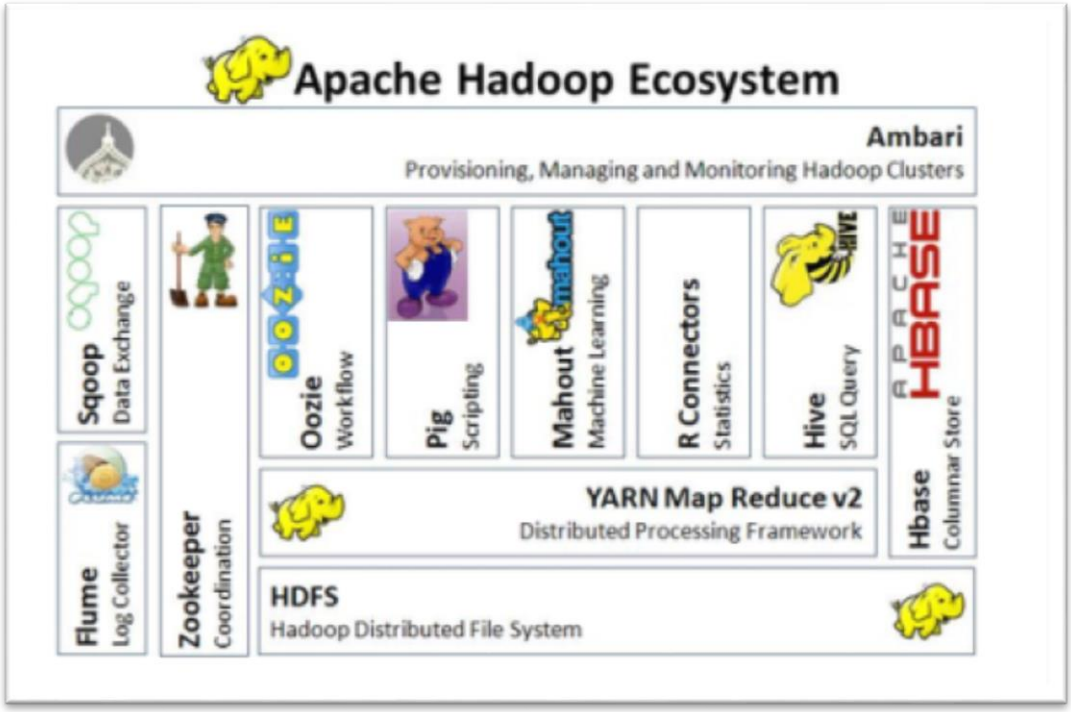
๑. Hadoop Common เป็นกลุ่มข้อมูล จำนวนมาก เพื่อการรองรับการทำงานของ Hadoop เช่น การกำหนดค่าหรือการปรับเปลี่ยนค่าข้อมูล

๒. Hadoop Distributed File System : HDFS ระบบแฟ้มข้อมูล เป็นการนำข้อมูลเข้า (Input Data) ที่มีขนาดใหญ่จำนวนมาก มาทำการแบ่งให้มีขนาดเล็กลง (Data Splitting) เพื่อกระจายไปประมวลผล

๓. Hadoop YARN ทำหน้าที่เป็นผู้บริหารทรัพยากร (Resource Management) ของแมปรีดิวซ์กำหนดและควบคุมการประมวลผล

๔. แมปรีดิวซ์ เป็นโปรแกรมที่ทำงานอยู่บน Hadoop ที่นำข้อมูลที่แบ่งเป็นข้อมูลเล็กๆแล้ว เข้าสู้ขั้นตอนแมปรีดิวซ์ โดยแมปรีดิวซ์ ประกอบด้วย ๒ ส่วน คือ ขั้นตอนแมป และขั้นตอนรีดิวซ์ เป็นการนำผลที่ได้

จากการประมวลผลแบบกระจายในขั้นตอนแมป ในแต่ละโหนดกลับมาทำการจัดเรียงลำดับและสรุปผลข้อมูล เพื่อแสดงผลลัพธ์การทำงานที่รวดเร็ว



รูปภาพ แสดงระบบการทำงานของ Apache Hadoop

ระบบซอฟต์แวร์ที่เกี่ยวข้องกับ Hadoop

HIVE ทำหน้าที่จัดการคลังข้อมูล (Data Warehouse) บนข้อมูล เพื่ออำนวยความสะดวกในการสอบถาม มีโครงสร้างแบบสอบถามข้อมูล (Query) การใช้ภาษาเหมือน SQL

PIG ทำหน้าที่อำนวยความสะดวกในการเขียนคำสั่งแบบภาษาสคริปต์ (Script) ที่ช่วยให้ประมวลผลข้อมูลโดยไม่ต้องเขียนแมปรีดิวซ์ด้วยภาษา JAVA ซึ่งจะทำให้คำสั่งเขียนกระชับและสั้น

SQOOP ทำหน้าที่จัดการเรื่องการถ่ายโอนข้อมูล

HBASE ทำหน้าที่จัดการเรื่องการอ่านและเขียนข้อมูลแบบเวลาจริง (Realtime) ได้ โดยจะนำข้อมูลมาเก็บในรูปแบบตารางใหญ่ คือไม่จำกัดจำนวนแถว หรือคอลัมน์

MAHOUT ทำหน้าที่จัดการเรื่องวิทยาศาสตร์ข้อมูล (Data Science) ที่นำข้อมูลใหญ่มาทำการวิเคราะห์และวิจัย และนำผลที่ได้มาจากการวิเคราะห์มาทำนายข้อมูล จากหลายๆ Algorithms

ZOOKEEPER ทำหน้าที่จัดการ บริการ ประสานงานเกี่ยวกับกระบวนการทำงานของแอปพลิเคชันแบบกระจาย และให้บริการทำซ้ำข้อมูลไปยังเครื่องไคลแอนด์และเซิร์ฟเวอร์ในระบบ

ประโยชน์ที่ได้รับ

๑. สามารถเข้าใจเกี่ยวกับความรู้พื้นฐานของข้อมูลขนาดใหญ่ (Big Data)

๒. มีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) ที่เกี่ยวกับข้อมูลขนาดใหญ่

๓. มีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ และสามารถนำไปประยุกต์ใช้กับการทำงานด้านการจัดการข้อมูล