ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคลากรภาครัฐทุกระดับ (BIG DATA)

โดยนางสาวสันศนีย์ ทองแย้ม
นักวิชาการแผนที่ภาพถ่ายชำนาญการ

วัตถุประสงค์

๑. เพื่อให้มีความรู้พื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data)

๒. เพื่อให้มีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่

๓. เพื่อให้มีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ

เนื้อหา

ในปัจจุบันการแข่งขันกันทางข้อมูลค่อนข้างสูง ใครมีข้อมูลมากกว่าย่อมได้เปรียบ ข้อมูลที่มากและหลากหลายย่อมก่อให้เกิดปัญหาในการใช้งาน จึงต้องมีการจัดการ จำแนกและวิเคราะห์ข้อมูล เพื่อให้องค์กรสามารถนำไปใช้งานได้อย่างมีประสิทธิภาพและทันเวลา รวมถึงการนำเสนอข้อมูลจำนวนมากนั้นต้องเข้าใจได้ง่ายและทันต่อการแข่งขัน

Big data คือข้อมูลขนาดใหญ่ มีทั้งโครงสร้างปกติและไม่มีโครงสร้าง ส่วนใหญ่ใช้ในเชิงธุรกิจทั้งหมด ประกอบด้วยคุณลักษณะ ๔ ประการ คือ

๑. Volum ข้อมูลมีขนาดใหญ่ มีปริมาณข้อมูลมาก ซึ่งสามารถเป็นได้ทั้งข้อมูลแบบ online หรือ offline

๒. Variety ข้อมูลมีความหลากหลาย เป็นได้ทั้งที่มีโครงสร้างและข้อมูลไม่สามารถจับ Pattern ได้

๓. Velocity ข้อมูลมีการเปลี่ยนแปลงตลอดเวลา มีการส่งผ่านข้อมูลอย่างต่อเนื่อง ทำให้การวิเคราะห์ข้อมูลแบบ Manual มีข้อจำกัด

๔. Veracity ข้อมูลมีตวามไม่ชัดเจน

ระดับของการวิเคราะห์ข้อมูล Big data ทำให้ข้อมูลที่เป็นข้อเท็จจริงซึ่งผ่านการวิเคราะห์อย่างเป็นระบบเพื่อใช้ประกอบการตัดสินใจ โดยระดับการวิเคราะห์ก็เป็นได้หลากหลาย แล้วแต่รูปแบบการนำไปใช้งาน

Descriptive Analytics เป็นการวิเคราะห์ในระดับที่บอกว่าเกิดอะไรขึ้น จำนวนเท่าไหร่ ถี่แค่ไหน เกิดเหตุการณ์สำคัญตอนไหน ตรงไหนบ้าง

Predictive Analytics เป็นการวิเคราะห์ในลักษณะที่ซับซ้อนขึ้นไปอีกขั้น คือ จะประเมินต่อไปว่าจะเกิดอะไรขึ้นต่อไป มีการให้ข้อมูลตัวชี้วัดของผลลัพธ์ที่อาจจะเกิดขึ้นถ้าแนวโน้มยังเป็นอย่างนี้ต่อไป

Prescriptive Analytics เป็นการวิเคราะห์ในลักษณะที่ซับซ้อนและยากที่สุด ไม่ใช่แค่พยากรณ์ผลที่จะเกิด แต่ยังให้คำแนะนำในทางเลือกต่างๆ และผลของทางเลือกมีผลดีและผลเสียอย่างไร เป็นการวิเคราะห์ที่เกี่ยวพันกับ Big data เป็นอย่างมาก

รูปแบบในการวิเคราะห์

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

Big data Analytics เป็นเครื่องมือที่มีความจำเป็นของภาครัฐในทุกประเทศ ในการขับเคลื่อนการบริหารราชการแผ่นดินอย่างมีประสิทธิภาพและขับเคลื่อนเศรษฐกิจให้ทันต่อการเปลี่ยนแปลงทางเทคโนโลยีอย่างก้าวกระโดดในศตวรรษที่ ๒๑ โดยมีแนวทาง ดังนี้

Hadoop เป็นโปรแกรมที่เข้ามาช่วยเราจัดการเก็บและประมวลผลข้อมูล อย่าง Big Data โดย Hadoop เป็นซอฟแวร์แบบโอเพนซอร์ช (Open Source) ของอาปาเชซอฟต์แวร์ ประกอบด้วยกลุ่มของชุดคำสั่งต่างๆ (Libraries) เพื่อช่วยอำนวยความสะดวกในการวิเคราะห์ข้อมูลขนาดใหญ่ๆ ได้อย่างมีประสิทธิภาพ

มีข้อมูล (Data) ใส่เข้าไปใน Hadoop (เครื่องมือ) ทำการวิเคราะห์ เกิดเป็นผลลัพธ์เอาไปใช้งาน

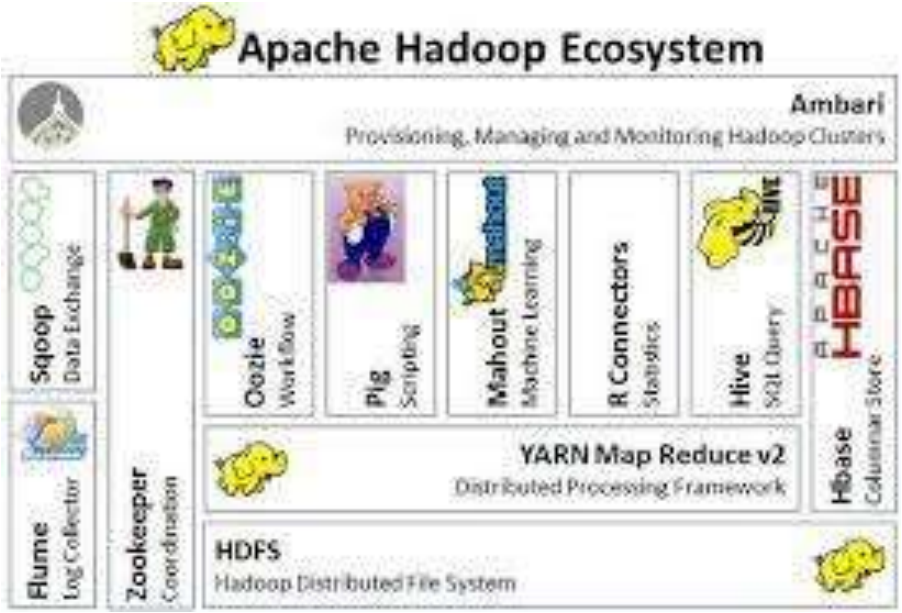
หลักการทำงานของ Hadoop ประกอบด้วย ๔ ส่วนหลัก

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) ของแมปรีดิวซ์ กำหนดและควบคุมการประมวลผล

๔. แมปรีดิวซ์ เป็นโปรแกรมที่ทำงานอยู่บน Hadoop ที่นำข้อมูลที่แบ่งเป็นข้อมูลเล็กๆแล้ว เข้าสู้ขั้นตอนแมปรีดิวซ์ โดยแมปรีดิวซ์ ประกอบด้วย ๒ ส่วน คือ ขั้นตอนแมป และขั้นตอนรีดิวซ์ เป็นการนำผลที่ได้จากการประมวลผลแบบกระจายในขั้นตอนแมป ในต่ละโหนดกลับมาทำการจัดเรียงลำดับและสรุปผลข้อมูล เพื่อแสดงผลลัพธ์การทำงานที่รวดเร็ว



ระบบที่เกี่ยวข้องกับ Hadoop

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

PIG เป็นซอฟแวร์ที่ทำหน้าที่อำนวยความสะดวกในการเขียนคำสั่งแบบภาษาสคริปต์ (Script) ที่ช่วยให้ประมวลผลข้อมูลโดยไม่ต้องเขียนแมปรีดิวซ์ด้วยภาษา JAVA ซึ่งจะทำให้คำสั่งเขียนกระชับและสั้น

SQOOP เป็นซอฟแวร์ที่ทำหน้าที่จัดการเรื่องการถ่ายโอนข้อมูล

HBASE เป็นซอฟแวร์ที่ทำหน้าที่จัดการเรื่องการอ่านและเขียนข้อมูลแบบเวลาจริง (Realtime) ได้ โดยจะนำข้อมูลมาเก็บในรูปแบบตารางใหญ่ คือไม่จำกัดจำนวนแถว หรือคอลัมน์

MAHOUT เป็นซอฟแวร์ที่ทำหน้าที่จัดการเรื่องวิทยาศาสตร์ข้อมูล (Data Science) ที่นำข้อมูลใหญ่มาทำการวิเคราะห์และวิจัย และนำผลที่ได้มาจากการวิเคราะห์มาทำนายข้อมูล จากหลายๆ Algorithms

ZOOKEEPER เป็นซอฟแวร์ที่ทำหน้าที่จัดการ บริการ ประสานงานเกี่ยวกับกระบวนการ ๔ ทำงานของแอพพลิเคชันแบบกระจาย และให้บริการทำซ้ำข้อมูลไปยังเครื่องไคลแอนด์และเซิร์เวอร์ในระบบ

ประโยชน์ที่ได้รับ

๑. มีความรู้พื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data)

๒. มีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่

๓. มีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อนำไปใประยุกต์ใช้กับการบริหารภาครัฐ