สรุปบทเรียนการเรียนรู้ผ่านสื่อการเรียนการสอนระบบ TGA E-learning
ประกอบตัวชี้วัดการพัฒนาความรู้ของบุคลากร รอบการประเมินที่ 1/2566

หลักสูตร ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการและบุคลากรภาครัฐทุกระดับ

สรุปโดยสาระสำคัญ มีดังนี้

1. **ความหมาย**

Big Data คือ ข้อมูลขนาดใหญ่ ตั้งแต่ 1 เพตะไบท์ (Petabyte) ขึ้นไป มีทั้งแบบโครงสร้างปกติและโครงสร้างข้อมูลที่ไม่มีรูปแบบ ซึ่งทั้งหมดเป็นข้อมูลที่ใช้ในเชิงธุรกิจ มักจะถูกนำไปใช้กับงานที่ต้องวิเคราะห์และมีความซับซ้อน เป็นต้น

2. **รูปแบบของข้อมูล Big Data** สามารถเป็นไปได้หลากหลาย ดังนี้

2.1 Behavioral Data ได้แก่ ข้อมูลเชิงพฤติกรรมการใช้งานต่างๆ เช่น Server Log พฤติกรรมการคลิกดูข้อมูล หรือข้อมูลการใช้ ATM เป็นต้น

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

2.3 Languages ตัวอย่างเช่น Text Message ข้อความที่ถูก Tweet เนื้อหาต่างๆ ในเว็บไซต์

2.4 Records ตัวอย่างเช่น ข้อมูลทางการแพทย์ ข้อมูลผลสำรวจที่มีขนาดใหญ่ ข้อมูลทางภาษี เป็นต้น

2.5 Sensors ตัวอย่างเช่น ข้อมูลอุณหภูมิ Accelerometer ข้อมูลทางภูมิศาสตร์ เป็นต้น

3. **คุณลักษณะของ Big Data** ประกอบด้วย 4 ประการ ดังนี้

3.1 Volume ข้อมูลมีขนาดใหญ่ มีปริมาณข้อมูลมาก ซึ่งสามารถเป็นได้ทั้งข้อมูลแบบ offline หรือ Online

3.2 Variety ข้อมูลมีความหลากหลาย สามารถเป็นได้ทั้งที่มีโครงสร้างและข้อมูลที่ไม่สามารถจับ Pattern ได้

3.3 Velocity ข้อมูลมีการเปลี่ยนแปลงตลอดเวลา อย่างรวดเร็ว มีการส่งผ่านข้อมูลอย่างต่อเนื่องในลักษณะ Streaming ทำให้การวิเคราะห์ข้อมูลแบบ Manual มีข้อจำกัด

3.4 Veracity ข้อมูลมีความไม่ชัดเจน (Untrusted Uncleaned)

4. **Data lake คือ** ที่เก็บส่วนกลางซึ่งช่วยให้จัดเก็บข้อมูลที่มีโครงสร้าง และไม่มีโครงสร้างในทุกขนาดได้ สามารถจัดเก็บข้อมูลตามที่เป็นโดยไม่ต้องวางโครงสร้าง และยังสามารถใช้การวิเคราะห์ประเภทต่างๆ ได้ ตั้งแต่ Dashboard และการแสดงภาพไปจนถึงการประมวลผล Big Data การวิเคราะห์แบบเรียลไทม์ และ Machine Learning เพื่อสร้างแนวทางการตัดสินใจที่ดีขึ้น **มีการแก้ไขข้อจำกัดหลายอย่างของ Data Warehouse ที่ใช้กันมานาน**

5. การวิเคราะห์ข้อมูล Big Data ทำให้มีข้อมูลที่เป็นข้อเท็จจริง ซึ่งผ่านการวิเคราะห์อย่างเป็นระบบ เพื่อใช้ประกอบการตัดสินใจ ระดับของการวิเคราะห์มีความหลากหลายขึ้นอยู่กับรูปแบบการนำไปใช้งาน ได้แก่

5.1 Descriptive Analytics เป็นการวิเคราะห์ในระดับที่บอกว่าเกิดอะไรขึ้น จำนวนเท่าไหร่ ถี่แค่ไหน เกิดเหตุการณ์สำคัญตอนไหน ตรงไหนบ้าง

5.2 Predictive Analytics เป็นการวิเคราะห์ในระดับที่ซับซ้อนขึ้นไปอีกขั้นหนึ่ง คือ เป็นการประเมินว่าจะเกิดอะไรขึ้นต่อไป มีการให้ข้อมูลตัวชี้วัดของผลลัพธ์ที่อาจจะเกิดขึ้น ถ้าแนวโน้มยังเป็นอย่างนี้ต่อไป

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

6. การใช้ประโยชน์จากข้อมูล Big Data ใช้เป็นข้อมูลวิเคราะห์ในเชิงทำนาย เช่น การพยากรณ์อากาศของกรมอุตุนิยมวิทยา การวิเคราะห์ทำนายพฤติกรรมของลูกค้าให้ตรงกับความต้องการของลูกค้า ข้อมูลสถิติทางการแพทย์ การจัดเตรียมหมอ ยารักษาโรค ให้เพียงพอกับผู้ป่วย การวิเคราะห์ทำนายโรคอุบัติใหม่ที่จะเกิดขึ้นในอนาคต เพื่อเตรียมการป้องกันและรับมือได้ทันสถานการณ์

7. กระบวนการจัดเก็บข้อมูล Big Data โดยมีกระบวนการหรือขั้นตอนในการจัดการข้อมูล แบ่งเป็น 3 ส่วน ได้แก่

7.1 แหล่งข้อมูล (Sources)

7.2 ใช้โปรแกรม Hadoop ประมวลผล วิเคราะห์ข้อมูลเชิงทำนาย

7.3 การนำผลวิเคราะห์ข้อมูลเชิงทำนาย ไปใช้ในรูปแบบแอปปลิเคชัน Applications เช่น การวิเคราะห์ข้อมูลทางธุรกิจ (Business Analytics) Custom Applications Packaged Applications เป็นต้น

8. กระบวนการทำงานของ Hadoop ประกอบด้วยระบบนิเวศของ Hadoop เรียกว่า Apache Hadoop Ecosystem มีทั้งหมด 11 โปรแกรม ตามภาพที่ 1 แต่ละโปรแกรมมีหน้าที่ในการทำงานแตกต่างกัน ไม่จำเป็นต้องติดตั้งโปรแกรมทั้งหมด สามารถเลือกติดตั้งเฉพาะโปรแกรมที่ต้องการใช้เฉพาะงานนั้นๆ ได้ ได้แก่

8.1 HDFS (Hadoop Distributed File System) ทำหน้าที่ตัดข้อมูลที่มีขนาดใหญ่ให้มีขนาดเล็กลงแล้วกระจายเอาไปประมวลผลการทำงาน

8.2 YARN Map Reduce v2 Distributed Processing Framework ทำหน้าที่ประสานการทำงานของ Hadoop ในการจัดการข้อมูลที่เสีย โดยการไปดึงข้อมูลอื่นๆ มาทดแทน

8.3 Hive SQL Query ทำหน้าที่เขียนคำสั่งหรือคำถามต่างๆ ของภาษา SQL ซึ่งเป็นภาษาที่จัดการเกี่ยวกับฐานข้อมูล โดยสามารถจัดการคำสั่งหรือคำถามเหล่านั้นส่งไปดึงข้อมูลจาก Hadoop มาวิเคราะห์เชิงทำนายตอบคำสั่งหรือคำถามนั้นๆ ได้

8.4 R Connectors Statistics ทำหน้าที่แสดงภาพข้อมูลที่เป็นสถิติ เป็น กราฟ แผนภูมิ เป็นต้น

8.5 Mahout Machine Learning ทำหน้าที่วิเคราะห์ข้อมูลในเชิงทำนายหรือพยากรณ์

8.6 Pig Scripting เป็นภาษาอุบัติใหม่ที่ใช้เขียนคำสั่งง่ายๆ สั้นๆ ในการใช้งานใน Hadoop

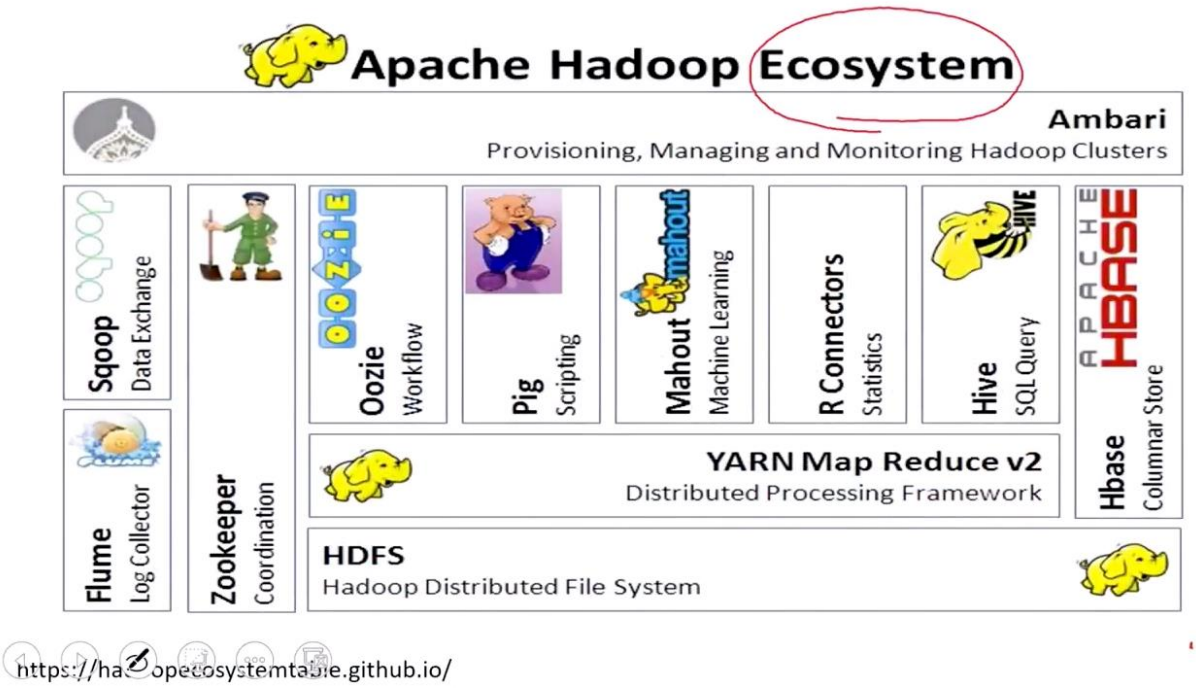
8.7 Oozie Workflow ทำหน้าที่จัดการลำดับขั้นตอนการทำงาน เช่น การทำนายหรือพยากรณ์ขั้นตอนในการสั่งซื้อสินค้า ตั้งแต่ใบสั่งซื้อสินค้า ส่งไปผลิตหรือยัง ผลิตแล้วถึงขั้นตอนไหน สุดท้ายได้อะไร เป็นต้น

8.8 Zookeeper Coordination ทำหน้าที่ควบคุมประสานระบบการทำงานทั้งหมดของ Hadoop ซ่อมแซมแก้ไขเครื่องมือที่มีปัญหา รันไม่ได้ ประมวลผลไม่ได้ ไฟฟ้าดับ ข้อมูลสูญหาย เพื่อให้ระบบสามารถทำงานต่อไปได้

8.9 Sqoop Data Exchange ทำหน้าที่ในการถ่ายโอนข้อมูล

8.10 Flume Log Collector ทำหน้าที่ประมวลผลข้อมูลที่มีการเคลื่อนที่ (Streaming data) ตลอดเวลาแบบเรียลไทม์

8.11 Hbase Columnar Store ทำหน้าที่เก็บข้อมูลที่มีขนาดใหญ่จาก Social network Facebook ซึ่งไม่สามารถเก็บในรูปแบบตารางได้ มาเก็บไว้ในรูปแบบของคอลัมน์เพียงอย่างเดียว



ภาพที่ 1 กระบวนการทำงานของ Hadoop

ผู้สรุปบทเรียน

นางสาวลดาวัลย์ นักฟ้อน

กลุ่มวิชาการเพื่อการพัฒนาที่ดิน สพข. 11

DGA Digital Government Development Agency

TDGA สถาบันพัฒนาบุคลากรภาครัฐด้านดิจิทัล THAILAND DIGITAL GOVERNMENT ACADEMY

ประกาศนียบัตร

ให้ไว้เพื่อแสดงว่า

ลดาวัลย์ นักฟ้อน

ได้ผ่านการอบรมด้วยระบบการเรียนออนไลน์ในบทเรียน
ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการ
และบุคลากรภาครัฐทุกระดับ

รวมระยะเวลาทั้งสิ้น 1 : 0 ชั่วโมง

โดยสถาบันพัฒนาบุคลากรภาครัฐด้านดิจิทัล
ภายใต้การดำเนินงานของสำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน)
ให้ไว้ ณ วันที่ 24 ม.ค. 2566

(นางไอรดา เหลืองวิไล)
รองผู้อำนวยการสำนักงานพัฒนารัฐบาลดิจิทัล
รักษาการแทนผู้อำนวยการสถาบันพัฒนาบุคลากรภาครัฐด้านดิจิทัล

d6232305