สรุปบทเรียนการพัฒนาความรู้

หลักสูตร ความรู้พื้นฐานเพื่อการวิเคราะห์ข้อมูลสำหรับข้าราชการ

และบุคลากรภาครัฐทุกระดับ

ชื่อ-สกุล นายสุรพงษ์ รังสิเสนา ณ อยุธยา **ตำแหน่ง** นักจัดการงานทั่วไปปฏิบัติการ

สังกัด ฝ่ายบริหารทั่วไป สำนักวิทยาศาสตร์เพื่อการพัฒนาที่ดิน

วันที่อบรม ๒๗ – ๒๘ มิถุนายน ๒๕๖๖

ผ่านระบบ e-Learning ของ สถาบันพัฒนาบุคลากรภาครัฐด้านดิจิทัล

วัตถุประสงค์ของหลักสูตร

๑. เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับข้อมูลขนาดใหญ่ (Big Data)

๒. เพื่อให้ผู้เรียนมีความรู้พื้นฐานเกี่ยวกับเครื่องมือวิเคราะห์ข้อมูล (Hadoop) เพื่อการทำงานเกี่ยวกับข้อมูลขนาดใหญ่

๓. เพื่อให้ผู้เรียนมีความเข้าใจพื้นฐานเกี่ยวกับการวิเคราะห์ข้อมูลขนาดใหญ่เพื่อการบริหารภาครัฐ

สรุปบทเรียน

จากเนื้อหาในหลักสูตรเป็นการเรียนรู้เกี่ยวกับข้อมูลขนาดใหญ่ที่เรียกกันว่า (Big Data) โดยในหลักสูตรเป็นการแนะนำการใช้เครื่องมือในการวิเคราะห์ข้อมูล หรือโปรแกรม Hadoop มาช่วยในการทำงานเกี่ยวกับข้อมูลขนาดใหญ่ เพื่อที่จะนำมาใช้สำหรับการบริหารภาครัฐ

๑) Big Data คืออะไร

Big Data คือ ข้อมูลที่มีขนาดใหญ่มาก ต้องเป็นข้อมูลที่มีขนาดมากกว่า ๑ เพตะไบต์ และเป็นข้อมูลที่มีความหลากหลาย รวมถึงข้อความ รูปภาพ และวีดีโอ เป็นต้น โดยสามารถแยกคุณลักษณะสำคัญของ Big Data ออกได้เป็น ๔ คุณลักษณะ ดังนี้

- Volume มีขนาดใหญ่มากกว่า ๑ เพตะไบต์
- Variety ต้องเป็นข้อมูลที่มีความหลากหลาย
- Velocity เป็นข้อมูลที่สามารถเพิ่มขึ้นได้อย่างรวดเร็ว
- Veracity ข้อมูลมีความถูกต้อง เป็นจริง ตรวจสอบได้

๒) จะนำ Big Data ไปใช้ประโยชน์ได้อย่างไร

เมื่อข้อมูลมีปริมาณมากจึงต้องอาศัยระบบประมวลผลที่มีประสิทธิภาพ สามารถรองรับปริมาณข้อมูลที่มีอยู่อย่างมหาศาล เราจะสามารถนำไปวิเคราะห์ข้อมูลในด้านต่างๆ เพื่อนำไปวางแผนและตัดสินใจ และนำไปใช้ในการบริหารภาครัฐได้ โดยในหลักสูตรเป็นการแนะนำในการใช้โปรแกรม Hadoop มาใช้ในการวิเคราะห์ข้อมูลของ Big Data โดยโปรแกรม Hadoop เป็นโปรแกรมที่สามารถหาใช้งานได้แบบฟรีไม่มีค่าใช้จ่าย สามารถดาวน์โหลดได้ง่าย โดยพิมพ์คำว่า “Apache Hadoop” ในหน้าค้นหาของ Google ได้เลย

๓) Hadoop คืออะไร มีกระบวนการทำงานอย่างไร

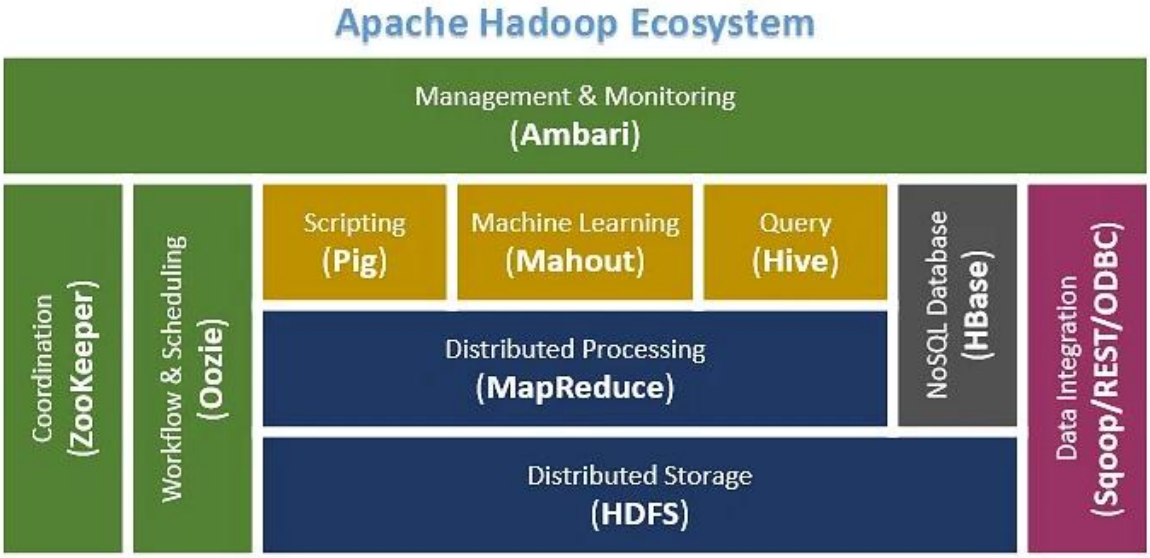
Hadoop หรือชื่อเรียกอย่างเป็นทางการคือ Apache Hadoop เป็น software framework สำหรับใช้ในการจัดการกับชุดข้อมูลที่มีขนาดใหญ่ และการประมวลผลแบบกระจาย เครื่องคอมพิวเตอร์ที่

เชื่อมต่อกันเป็นกลุ่ม เนื่องจากทุก ๆ วัน จะมีข้อมูลหลั่งไหลเข้ามารวมกันเป็นจำนวนมาก ทำให้ไม่สามารถใช้หน่วยเก็บข้อมูล และหน่วยประมวลแบบธรรมดาอันเดียวได้อีก เพราะนอกจากข้อมูลที่เข้ามามากมายมหาศาลแล้ว ข้อมูลยังเข้ามาหลายรูปแบบ ไม่ว่าจะเป็น ข้อความ รูปภาพ วีดีโอ และเสียง จะมีโครงสร้างข้อมูล (Data Structure) ที่แตกต่างกัน ทำให้เกินกำลังคอมพิวเตอร์หนึ่งเครื่องจะสามารถรับไหวอีก ดังนั้น Hadoop จึงเข้ามามีบทบาทในการแก้ไขเรื่องนี้

โดย Hadoop จะมีหน้าที่หลัก ๆ อยู่ ๒ อย่าง

- บริหารจัดการหน่วยเก็บข้อมูล (storage management)
- บริหารจัดการทรัพยากรสำหรับการประมวลผล (computing resource management)

ซึ่งใน Hadoop ก็จะมี software หลายตัวที่ถูกพัฒนาขึ้นมาเป็น Ecosystem ของ Hadoop โดย software เหล่านี้แต่ละตัวก็จะดึงความสามารถเด่น ๆ แต่ละด้านของ Hadoop ออกมาเพื่อให้เกิดประสิทธิภาพในการทำงาน ให้เราสามารถเลือกใช้งานได้ตามความเหมาะสมกับลักษณะงาน ดังที่แสดงในตัวอย่างด้านล่าง



- **Apache Flume** : เครื่องมือที่ใช้เก็บ จัดเรียง และส่งข้อมูลให้กับ HDFS
- **Apache HBase** : เป็น database แบบ nonrelational
- **Apache Hive** : คือ data warehouse ที่ทำการเก็บข้อมูลเพื่อ query และวิเคราะห์ข้อมูลที่ต้องการ
- **Cloudera Impala** : software ส่วนที่จัดการทำงานแบบ parallel ให้บน Hadoop
- **Apache Oozie** : ไว้จัดการ workflow รวมถึงตั้ง job การทำงานบน Hadoop
- **Apache Phoenix** : ใช้สำหรับ connection ไปยัง HBase และใช้งาน SQL query
- **Apache Pig** : platform ที่ใช้สร้าง program เพื่อทำงานบน Hadoop
- **Apache Sqoop** : เครื่องมือสำหรับส่งข้อมูลขนาดใหญ่ระหว่าง Hadoop และส่วนของข้อมูลที่มีโครงสร้าง หรือ relational database
- **Apache Spark** : engine สำหรับ Big data ที่ใช้ประมวลผลข้อมูลแบบ streaming และยังรองรับ SQL ด้วย
- **Apache Storm**: เป็นตัวส่งต่อข้อมูลแบบ real-time
- **Apache ZooKeeper**: เหมือนเป็น server ที่ใช้เก็บข้อมูลทีจะถูกส่งต่อให้อีกที

ประโยชน์ที่ได้รับต่อตนเอง

สามารถนำความรู้และเทคนิคที่ได้นำมาปรับใช้กับการทำงานในการวิเคราะห์ข้อมูลขนาดใหญ่ เพื่อดึงข้อมูลที่สำคัญมาใช้ให้ถูกต้องและเหมาะสม เพื่อเป็นการเพิ่มศักยภาพการทำงานของตัวเองได้อย่างดีและมีประสิทธิภาพ

ประโยชน์ที่ได้รับต่อหน่วยงาน

ช่วยให้การขับเคลื่อนงานของฝ่ายสามารถเป็นไปได้อย่างมีประสิทธิภาพ รวดเร็ว และตอบสนองกับนโยบายของสำนักฯ และของกรมพัฒนาที่ดิน