# สรุปแนวทางการเตรียมตัวอ่านหนังสือ ภาคการศึกษา พิเศษ 64 และ 1/65

1. **99703 การจัดการเทคโนโลยีสารสนเทศและการสื่อสารเชิงกลยุทธ์**
   - ความสอดคล้องเชิงกลยุทธ์เทคโนโลยีสารสนเทศ โดยโจทย์ให้ยกตัวอย่างธุรกิจ ซึ่งนักศึกษาจะต้องวิเคราะห์มุมมองความสอดคล้องเชิงกลยุทธ์ ตามองค์ประกอบต่างๆ ของแบบจำลองความสอดคล้องเชิงกลยุทธ์ และประเมินระดับวุฒิภาวะความสอดคล้องเชิงกลยุทธ์ตามมิติและลักษณะของแต่ละระดับวุฒิภาวะ พร้อมกับวาดภาพประกอบการเขียนตอบ

2. **ชุดวิชา 99705 ความมั่นคงด้านเทคโนโลยีสารสนเทศและการสื่อสาร**
   - ศึกษากรณีศึกษาต่างๆ งานวิจัยที่เกี่ยวกับภัยคุกคามต่างๆ ที่เกิดขึ้น การถูกละเมิดความเป็นส่วนบุคคลของข้อมูล Social Engineering, Phishing, Spam mail หรืออื่นๆ
   - วิเคราะห์ช่องโหว่ ภัยคุกคาม และผลกระทบที่เกิดขึ้น
   - แนวทางการป้องกันโดยใช้เทคโนโลยีหรือซอฟต์แวร์ที่เกี่ยวข้อง

3. **99708 ระเบียบวิธีวิจัยและเครื่องมือในการพัฒนาระบบด้านเทคโนโลยี**
   - การเขียนแผนภาพยูสเคส (Use Case Diagram) จากกรณีศึกษา
   - การสร้างแผนภาพกิจกรรม (Activity Diagram) จากกรณีศึกษา
   - การสร้างแผนภาพคลาส (Class Diagram) จากกรณีศึกษา

4. **ชุดวิชา 99710 เทคโนโลยีเคลื่อนที่ไร้สายและการประยุกต์**
   - ศึกษาตัวอย่างจากกรณีศึกษาในกิจกรรมสัมมนาเข้ม / สัมนาเสริม เกี่ยวกับการประยุกต์บูรณาการ และรวบรวมข้อมูลจากไอโอทีเซนเซอร์ต่างๆ นำมาบันทึก วิเคราะห์ และประมวลผลต่างๆ (IOT Data Management / Processing / Cloud)

5. **ข้อสอบบูรณาการ**
   - บูรณการหลักการของเทคโนโลยีสารสนเทศและการสื่อสาร รวมถึงเทคโนโลยีดิจิทัลต่างๆ กับชุดวิชาในหลักสูตรวิทยาศาสตรมหาบัณฑิต แขนงเทคโนโลยีสารสนเทศและการสื่อสาร

**สรุปประเด็นวิชา 99707 ระบบสารสนเทศภูมิศาสตร์และการประยุกต์**

แนวข้อสอบคือ

1. ทำการศึกษาเกี่ยวกับ แนวทางการประยุกต์ใช้งานของระบบสารสนเทศภูมิศาสตร์ ในด้านต่าง ๆ ว่ามีอะไรบ้าง
2. ทำการศึกษาเกี่ยวกับ ประเภทและลักษณะข้อมูล รวมถึงกระบวนการทำงานในระบบสารสนเทศภูมิศาสตร์ ว่าเป็นอย่างไร

**สรุปประเด็นวิชา 99709 ธุรกิจดิทัลและการประยุกต์**

ธุรกิจดิจิทัลเป็นการสร้างสรรของธุรกิจใหม่ที่ออกแบบโดยการทำให้ภาพของโลกดิจิทัลและโลกทางกายภาพเลือนหายเข้าด้วยกัน เป็นการผสมผสานระหว่างเทคโนโลยีดิจิทัลและการทำงานของคนที่ทำให้เกิดสินค้าและบริการรูปแบบใหม่ ขณะที่งานด้านการตลาด การซื้อขาย การชำระเงิน หรือบริการหลังการขายอื่นๆ จะเกิดขึ้นในโลกดิจิทัลแทน ธุรกิจได้เปลี่ยนแนวทางจากยุคอะนาล็อก (Analog) มาเป็น Web base แล้วเปลี่ยนเป็น E-Commerce/E-Business แล้วจึงมาเป็น Digital marketing จนมาเป็นยุคธุรกิจดิจิทัลในที่สุด

ธุรกิจดิจิทัลต้องสร้างกลยุทธ์ที่ให้ความสำคัญกับการสร้างคุณค่า (Value Creation) ด้วยการเสนอการบริการที่หลากหลายให้ลูกค้าเลือกใช้ตามอัธยาศัย เพื่อให้มั่นใจว่าจะได้คุณค่าจริงในระหว่างการใช้หรือหลังการใช้สินค้าและบริการที่ได้ซื้อไป ธุรกิจดิจิทัลให้ความสำคัญกับการแก้ปัญหาและตอบโจทย์ของลูกค้าเฉพาะตัวให้มากที่สุด รูปแบบของธุรกิจดิจิทัลที่เน้นการสร้างคุณค่านั้นเป็นรูปแบบธุรกิจดิจิทัลที่ยั่งยืนกว่า แต่ทั้งนี้อยู่บนสมมุติฐานว่าธุรกิจต้องมีความสามารถในด้านเทคโนโลยีดิจิทัล (Digital Capability) ที่จำเป็นต่อการทำธุรกิจ

เทคโนโลยีแพลตฟอร์ม (Technology Platform) คือ เทคโนโลยีสารสนเทศและการสื่อสารที่สนับสนุนการทำธุรกิจดิจิทัล ซึ่งประกอบด้วยส่วนสำคัญหลัก 5 ส่วน ได้แก่

1 สื่อสังคม (Social) *สื่อสังคม (Social Media)* หรือ เทคโนโลยีสื่อสังคม (Social technology) คือสื่ออิเล็กทรอนิกส์ ซึ่งเป็นสื่อกลางที่ให้บุคคลทั่วไปมีส่วนร่วมสร้างและแลกเปลี่ยนความคิดเห็นต่างๆ ผ่านอินเทอร์เน็ตได้

2 โมบาย (Mobile) คือ อุปกรณ์สื่อสารไร้สายที่เคลื่อนที่ได้ ซึ่งผู้ใช้สามารถในการติดต่อสื่อสารพูดคุยแบบโทรศัพท์ได้ และเข้าถึง application ต่างๆ บนอุปกรณ์สื่อสารไร้สายเคลื่อนที่ได้ในรูปแบบเรียลไทม์ (real time)

3. การวิเคราะห์ข้อมูลขนาดใหญ่ ( Big data Analytics) คือ การใช้เครื่องมือวิเคราะห์ข้อมูลจำนวนมหาศาลในเชิงลึก

4. การประมวลผลกลุ่มเมฆ (Cloud Computing) คือบริการที่ครอบคลุมถึงการให้ใช้กำลังประมวลผล หน่วยจัดเก็บข้อมูล และระบบออนไลน์ต่างๆ จากผู้ให้บริการ เพื่อลดความยุ่งยากในการติดตั้ง ดูแลระบบ ช่วยประหยัดเวลา และลดต้นทุนในการสร้างระบบคอมพิวเตอร์และเครือข่ายเอง

5. อินเทอร์เน็ตของสรรพสิ่ง (Internet of Things) คือ สภาพแวดล้อมอันประกอบด้วยสรรพสิ่งที่สามารถสื่อสารและเชื่อมต่อกันได้ผ่านโพรโทคอลการสื่อสารทั้งแบบใช้สายและไร้สาย โดยสรรพสิ่งต่าง ๆ มีวิธีการระบุตัวตนได้ รับรู้บริบทของสภาพแวดล้อมได้ และมีปฏิสัมพันธ์โต้ตอบและทำงานร่วมกันได้ **ขอให้นักศึกษาไปศึกษารายละเอียดที่เกี่ยวข้องในส่วนนี้**

กระบวนการสร้างนวัตกรรมสำหรับธุรกิจดิจิทัลประกอบด้วย 6 ส่วน คือ 1) การจัดการประสบการณ์ลูกค้า การวิเคราะห์ลูกค้าตั้งแต่จุดแรกที่ลูกค้าได้พบเห็นและประสบการณ์ที่เกิดจากการนำเสนอ 2) นวัตกรรมทางสินค้าและบริการ เป็นการสร้างสินค้าและบริการใหม่ หรือพัฒนา และปรับปรุงสินค้าที่มีอยู่ 3) เทคโนโลยีดิจิทัล เป็นการนำเทคโนโลยีเข้ามาสร้างนวัตกรรม 4) การปฏิบัติตามดิจิทัล เป็นการจัดการงานที่ต้องทำในแต่ละวันรวมถึงการสนับสนุนการให้บริการต่าง ๆ สำเร็จลุล่วง 5) การจัดการความเสี่ยง เป็นกระบวนการทำงานที่ช่วยให้ IT Management ประสบความสำเร็จ การวางมาตรการในการป้องกันความเสี่ยงต่าง ๆ 6) การปรับปรุงการควบคุมดูแลกิจการ การกำกับดูแลกิจการที่ดี แสดงให้เห็นว่ามีระบบบริหารจัดการที่มีประสิทธิภาพ โปร่งใส ตรวจสอบได้ **ขอให้นักศึกษาไปศึกษารายละเอียดที่เกี่ยวข้องในส่วนนี้**

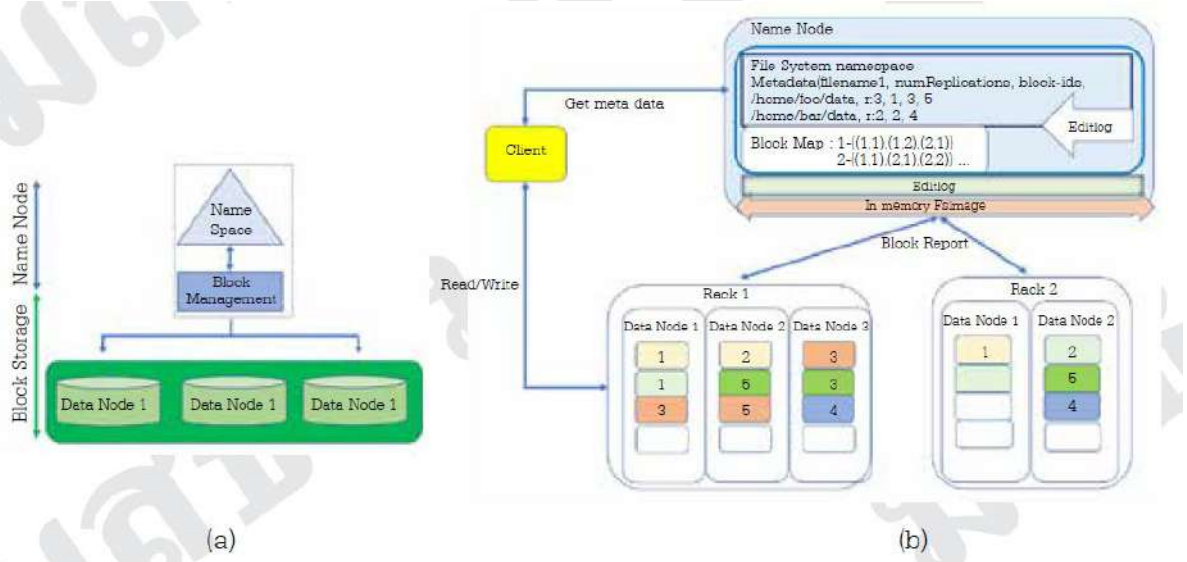
--99711--

**แนวทางการทบทวนเบื้องต้นเกี่ยวกับการสอบประมวล**

**1. สถาปัตยกรรมของเอชดีเอฟเอสและหลักการในการจัดการแฟ้มข้อมูล**

มีแนวคิดประกอบด้วย

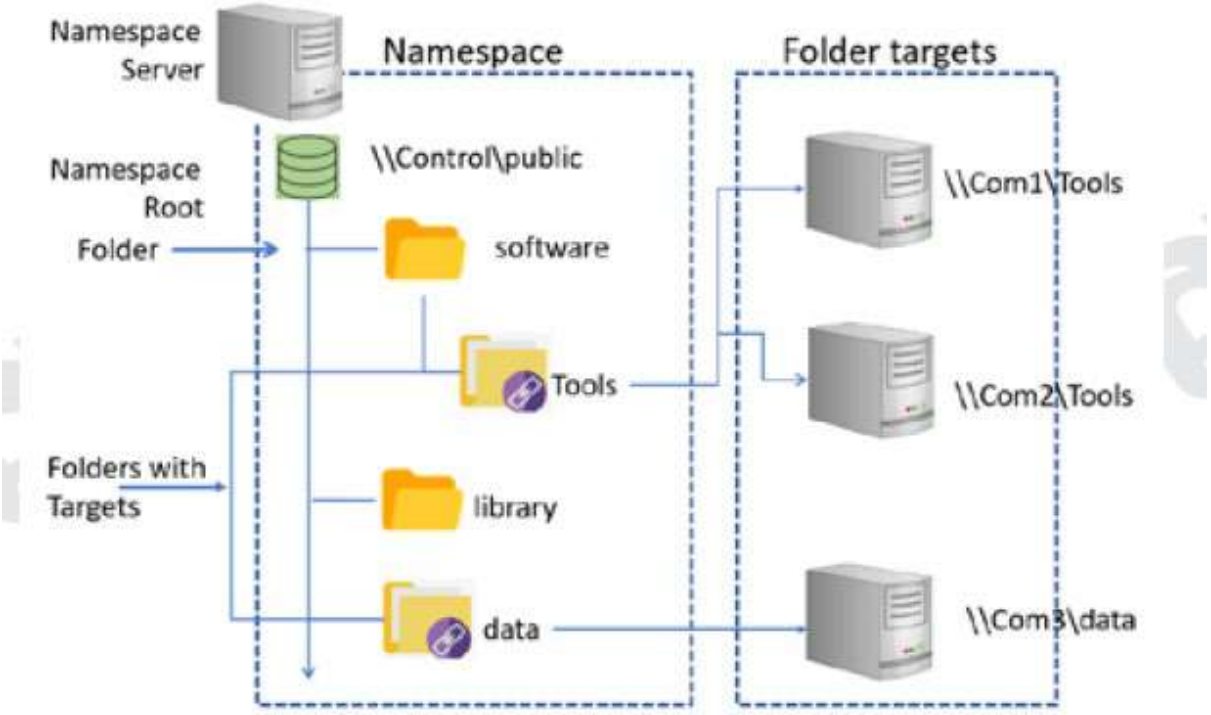
**1.1** สถาปัตยกรรมเอชดีเอฟเอส ใช้แนวคิดพื้นฐานการประมวลผลแบบมาสเตอร์/สเลฟ มีเนมโหนดเป็นมาสเตอร์ และดาตาโหนดจำนวนหนึ่งเป็นสเลฟ เนมโหนดมีหน้าที่หลักในการจัดการไฟล์ซีสเต็ม และจัดการเก็บข้อมูลของไฟล์ในการอ่านและเขียน โดยเนมโหนดต้องทำการแมปชื่อแฟ้มข้อมูลในมุมของผู้ใช้ไปยังที่เก็บจริงในดาตาโหนด และดาตาโหนดมีหน้าที่ในการจัดเก็บข้อมูลจริง ๆ โดยมีการจัดทำสำเนาเพื่อให้ระบบมีสภาพความคงทนพร้อมใช้งาน




**ภาพที่ 5.8 องค์ประกอบของ HDFS และการจัดเก็บไฟล์**

*** หมายเหตุ ให้ศึกษาคำอธิบายเพิ่มเติมในเอกสารการสอนของภาพที่ 5.8 โดยควรสามารถอธิบายถึงการทำงานของแต่ละองค์ประกอบในภาพและโดยรวมได้

1.2 หลักการในการจัดการแฟ้มข้อมูลแบบกระจายของเนมโหนด ในการจัดการไฟล์ซีสเต็มเนมสเปซใช้หลักการเดียวกันกับระบบปฏิบัติการทั่วไป การจัดการไฟล์ซีสเต็มเนมสเปซประกอบด้วย แฟ้มข้อมูล 2 แฟ้ม คือ Fslmage และ Editlog แฟ้ม Fslmage จะเก็บการแมบชื่อแฟ้มข้อมูลไปยังที่จัดเก็บจริงในดาตาโหนด และส่วนใหญ่จะเก็บในหน่วยความจำ การเปลี่ยนแปลงแฟ้มข้อมูลจะบันทึกใน Editlog ก่อน แล้วจึงนำมาปรับปรุงใน FsImage แนวคิดในการปรับปรุงใช้หลักการของเช็คพอยท์ นอกจากนี้เนมโหนดยังมีหน้าที่ในการจัดสมดุลของจำนวนสำเนา รวมทั้งการเลือกที่จัดวางสำเนาและการสมดุลในแต่ละแร็ค เพื่อให้รองรับสภาพความคงทนในการประมวลผล




**ภาพที่ 5.9 ตัวอย่างการจัดโครงสร้างและการแมปที่เก็บจริง**

*** หมายเหตุ ให้ศึกษาคำอธิบายเพิ่มเติมในเอกสารการสอนของภาพที่ 5.9 – 5.10 โดยควรสามารถอธิบายถึงการทำงานของแต่ละองค์ประกอบในภาพและโดยรวมได้

ทุกครั้งที่มีการปรับปรุงไฟล์ซิสเต็มเนมสเปซ[2] เนมโหนด โดยใช้ทรานแซกชันล็อก (Transaction log) ที่เรียกว่า EditLog เพื่อบันทึกเก็บข้อมูลอย่างถาวร ทุกครั้งที่มีการเปลี่ยนแปลงเมทาดาตาของไฟล์ซิสเต็ม (File System Metadata) ตัวอย่างเช่น ไคลเอ็นต์ต้องการสร้างไฟล์ใหม่ใน HDFS ไคลเอ็นต์ส่งรายละเอียดชื่อแฟ้มข้อมูล ขนาดของแฟ้มข้อมูล ให้เนมโหนด และเนมโหนดต้องเพิ่มรายการใน EditLog ทำนองเดียวกันกับการเปลี่ยนแปลงจำนวนสำเนา ก็บันทึกรายการใหม่ใน EditLog โครงสร้างแฟ้มข้อมูลของ Editlog เนมโหนดใช้โครงสร้างแฟ้มข้อมูลที่เก็บโดยระบบปฏิบัติการพื้นฐานทั่วไปในการเก็บ Editlog

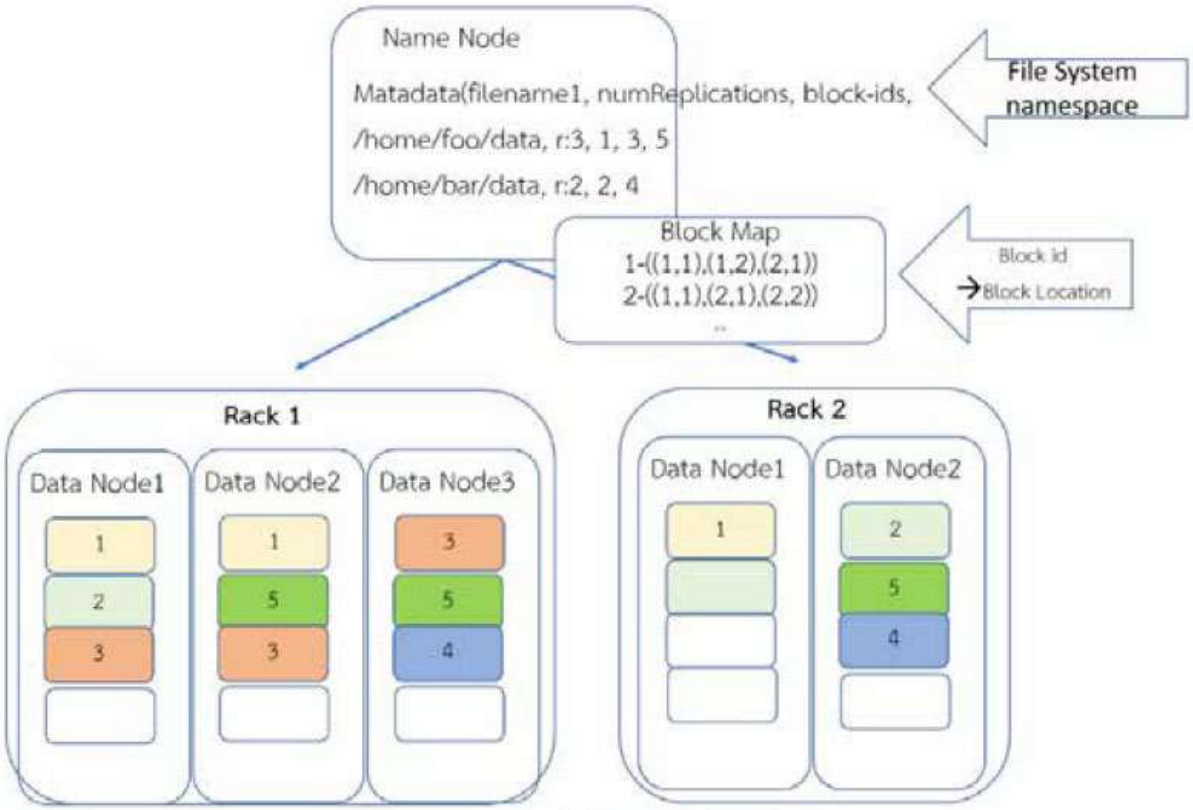
แนวทางในการนำรายการปรับปรุงที่ถูกบันทึกเก็บใน Editlog มาปรับปรุง FsImage คือ ใช้หลักการของจุดตรวจสอบ หรือเช็คพอยต์ (checkpoint) โดยมีการกำหนดค่าระดับของเช็คพอยท์ไว้ (Configuration Threshold) เมื่อมีการเริ่มต้นประมวลผลเนมโหนด หรือเกิดสัญญาณแจ้งเตือนจากระบบเช็คพอยท์ เนมโหนดอ่าน FsImage และ EditLog จากดิสก์ แล้วปรับปรุง FsImage ในหน่วยความจำใหม่โดยการอ่านข้อมูลทั้งหมดใน EditLog และบันทึก FsImage รุ่นใหม่ลงดิสก์ และลบรายการใน EditLog สำหรับรายการที่ได้ปรับปรุงไปแล้ว

การประมวลผลจุดเช็คพอยต์ก็เพื่อให้ HDFS มีข้อมูลของ File System Metadata ที่ถูกต้อง โดยหลักในการทำงานดังกล่าวถือว่ามีประสิทธิภาพ แม้ว่าต้องเสียเวลาในการอ่าน FsImage ทั้งหมด แล้วจึงใช้ข้อมูล EditLog ปรับปรุงในหน่วยความจำ แล้วบันทึกกลับในดิสก์แทนการปรับปรุง โดยการบันทึก FsImage กระบวนการจุดเช็คพอยต์มีกำหนดเวลาในปรับปรุง หรือตามจำนวนรายการที่บันทึกใน EditLog ถ้าตัวแปรใดถึงค่าที่กำหนดกระบวนการตรวจสอบจะเกิดขึ้น

ดาตาโหนดเก็บข้อมูล HDFS ในแฟ้มข้อมูล ซึ่งเป็นแฟ้มข้อมูลท้องถิ่นของระบบแฟ้มข้อมูล ดาตาโหนดไม่รายละเอียดข้อมูลเกี่ยวกับแฟ้มข้อมูลของ HDFS ระบบจัดเก็บบล็อกข้อมูลของ HDFS แยกเป็นแต่ละแฟ้มในระบบแฟ้มข้อมูลท้องถิ่น ดาตาโหนดไม่ได้สร้างทุกไฟล์ในไดเร็กทอรีเดียวกัน ระบบใช้วิธีฮิวริสติก[3] ในการพิจารณาเพื่อให้เกิดประสิทธิภาพสูงสุดในการพิจารณาจำนวนไฟล์ในหนึ่งไดเร็กทอรี และจำนวนไดเร็กทอรีย่อยที่เหมาะสม เป็นการไม่เหมาะในการสร้างแฟ้มข้อมูลท้องถิ่นทุกแฟ้มข้อมูลในไดเร็กทอรีเดียวกัน และอาจไม่มีประสิทธิภาพในการสร้างแฟ้มข้อมูลขนาดใหญ่ในหนึ่งไดเร็กทอรี

เมื่อดาตาโหนดเริ่มต้นทำงาน จะสำรวจแฟ้มข้อมูลท้องถิ่นทั้งหมด และสร้างรายการของดาตาบล็อกที่สอดคล้องกับแฟ้มข้อมูลท้องถิ่น และส่งข้อมูล หรือเป็นรายงานไปยังเนมโหนด ซึ่งรายงานนี้เรียกว่าบล็อกรีพอร์ต (Blockreport)

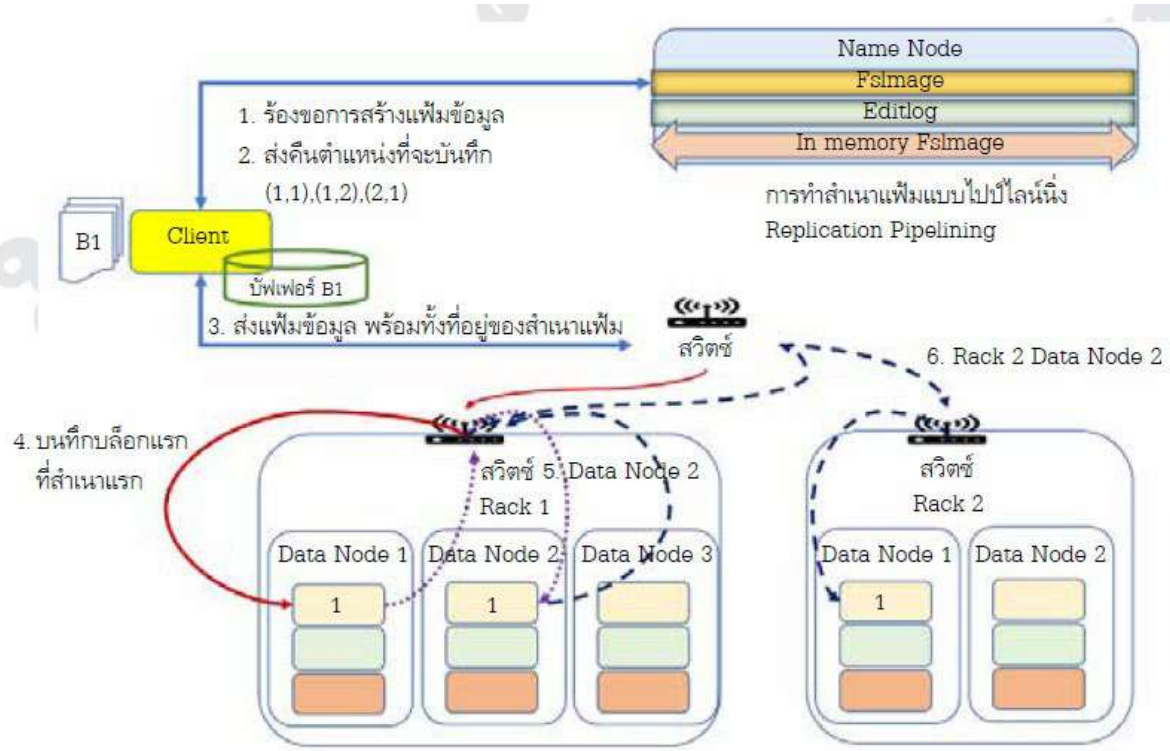
การทำสำเนาข้อมูล (Data Replication) ด้วยวัตถุประสงค์ของฮาดูปที่ต้องรองรับแฟ้มข้อมูลที่มีข้อมูลปริมาณมากได้ การออกแบบจึงทำการแบ่งแฟ้มข้อมูลขนาดใหญ่ เป็นบล็อกที่มีขนาดเท่ากัน และบันทึกเก็บในหลายดาตาโหนด และออกแบบระบบไฟล์ซิสเต็มเนมสเปซเพื่อช่วยในการแมปบล็อกข้อมูลกับที่จัดเก็บจริง อย่างไรก็ตามการเก็บข้อมูลที่ดาตาโหนดซึ่งอยู่ระยะไกล (อาจอยู่คนละสถานที่ หรือต้องเชื่อมต่อผ่านเครือข่ายสื่อสาร เช่น สวิชต์) อาจพบปัญหา เช่น ไม่สามารถติดต่อเครื่องคอมพิวเตอร์ เพราะระบบสื่อสารล่ม หรือเครื่องคอมพิวเตอร์เกิดเสียหายจากฮาร์ดแวร์ไม่ทำงาน




**ภาพที่ 5.10 การจัดเก็บสำเนาในดาตาโหนด**

*** หมายเหตุ ให้ศึกษาคำอธิบายเพิ่มเติมเกี่ยวกับหลักการสำเนาข้อมูลและตัวอย่างในการจัดวางสำเนาของภาพที่ 5.9 – 5.10 โดยควรสามารถอธิบายถึงการทำงานของแต่ละองค์ประกอบในภาพและโดยรวมได้

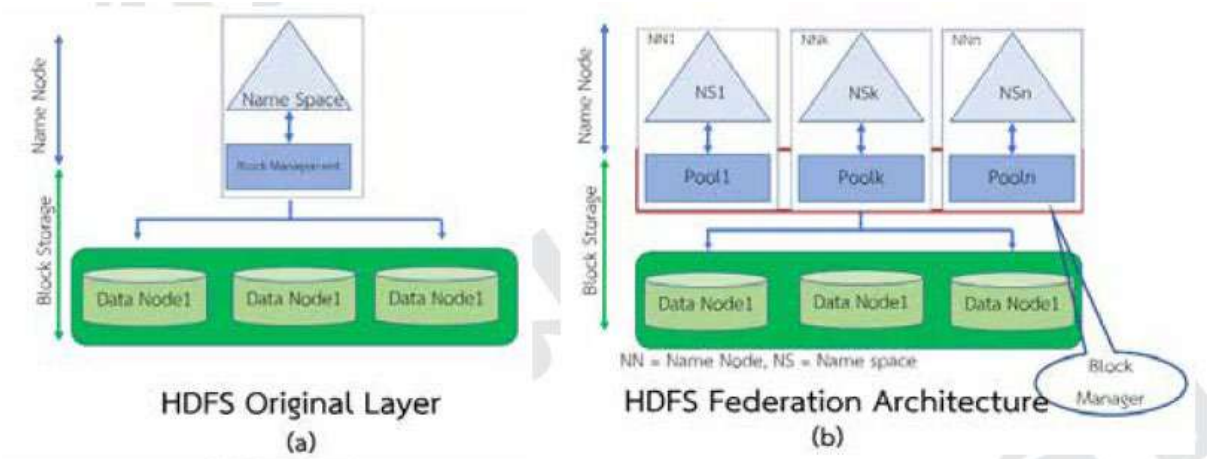
**1.3** การจัดการข้อมูลในดาตาโหนด ข้อมูลจริงถูกบันทึกในดาตาโหนด โดยแบ่งออกเป็นบล็อก ๆ ละ 128 MB และต้องทำการบันทึกสำเนาตามข้อมูลที่กำหนดมาจากเนมโหนด โดยหลักการในการบันทึกสำเนาเรียกว่า ไปป์ไลน์ เพื่อลดเวลาในการติดต่อสื่อสารระหว่างไคลเอ็นต์และเนมโหนด




**ภาพที่ 5.12 การบันทึกแฟ้มข้อมูลของดาตาโหนดแบบไปป์ไลน์**

*** หมายเหตุ ให้ศึกษาคำอธิบายเพิ่มเติมในเอกสารการสอนของภาพที่ 5.11 – 5.12 โดยควรสามารถอธิบายถึงการทำงานของแต่ละองค์ประกอบในภาพและโดยรวมได้

**1.4** การจัดการสภาพคงทนในการประมวลผล และรองรับการขยายของปริมาณข้อมูล การจัดการปริมาณข้อมูลขนาดใหญ่ใช้หลักการแบ่งแฟ้มข้อมูลเป็นแฟ้มย่อย ส่วนสภาพความคงทนมีทั้งส่วนของดาตาโหนดและเนมโหนด ซึ่งดาตาโหนดไม่มีข้อจำกัดในการจัดเก็บแฟ้มข้อมูล และมีการจัดทำสำเนาที่คาดว่าเพียงพอต่อความล้มเหลวในการทำงาน ขณะเดียวกันก็มีการจัดสมดุลจำนวนสำเนา และการจัดวางที่เหมาะสม สำหรับเนมโหนดใช้ 2 หลักการ คือ มีเนมโหนดสำรอง และใช้ เอชดีเอฟเอส เฟเดอเรชัน แบบแรกจะรองรับเฉพาะการล้มเหลวในการทำงาน และไม่รองรับการขยายของแฟ้มข้อมูล ทั้งนี้เพราะ FsImage จัดเก็บในหน่วยความจำเอชดีเอฟเอส จึงแก้ปัญหาการเพิ่มขึ้นของปริมาณข้อมูล และจำนวนแฟ้มข้อมูลได้




**ภาพที่ 5.13 HDFS Federation และ HDFS แบบดั้งเดิม**

*** หมายเหตุ ให้ศึกษาคำอธิบายเพิ่มเติมในเอกสารการสอนของภาพที่ 5.13 โดยควรสามารถอธิบายถึงการทำงานของแต่ละองค์ประกอบในภาพและโดยรวมได้

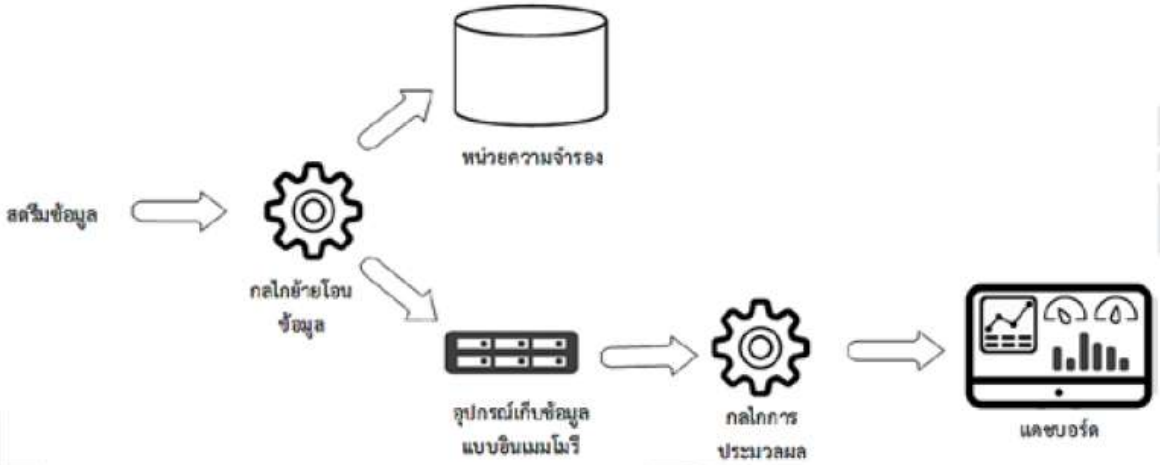
## 2. เครื่องมือในการวิเคราะห์ข้อมูลขนาดใหญ่

มีแนวคิดประกอบด้วย

**2.1** การนำข้อมูลขนาดใหญ่มาวิเคราะห์เพื่อใช้ประโยชน์ทั้งในแง่การตีความหมายข้อมูล การวิเคราะห์หาความสัมพันธ์ รูปแบบ หรือค้นหาความรู้ที่ซ่อนอยู่ในข้อมูลปริมาณมหาศาล นอกจากต้องมีความรู้ความเข้าใจในแนวคิดสำคัญของเทคนิคที่จะนำมาใช้งาน ยังต้องเลือกใช้เครื่องมือทางเทคโนโลยีสารสนเทศให้ถูกต้องเหมาะสม พิจารณาความสอดคล้องกับลักษณะข้อมูล ความเข้ากันได้กับเทคโนโลยีพื้นฐานที่แต่ละองค์กรใช้อยู่ รวมทั้งสามารถรองรับความต้องการในการนำไปใช้ประโยชน์แต่ละด้าน

ในข้อนี้มีประเด็นที่กล่าวถึงได้แก่

**2.1.1** รูปแบบการทำงานและการประมวลผลข้อมูล ที่มักใช้ในการวิเคราะห์ข้อมูลขนาดใหญ่ ได้แก่ การวิเคราะห์แบบกลุ่ม การสอบถามข้อมูลเชิงโต้ตอบ การวิเคราะห์แบบเรียลไทม์ การประมวลผลเหตุการณ์แบบซับซ้อน การใช้โมเดลในการวิเคราะห์เชิงพยากรณ์ และการวิเคราะห์ข้อมูลในระบบแนะนำแบบเรียลไทม์ เป็นต้น

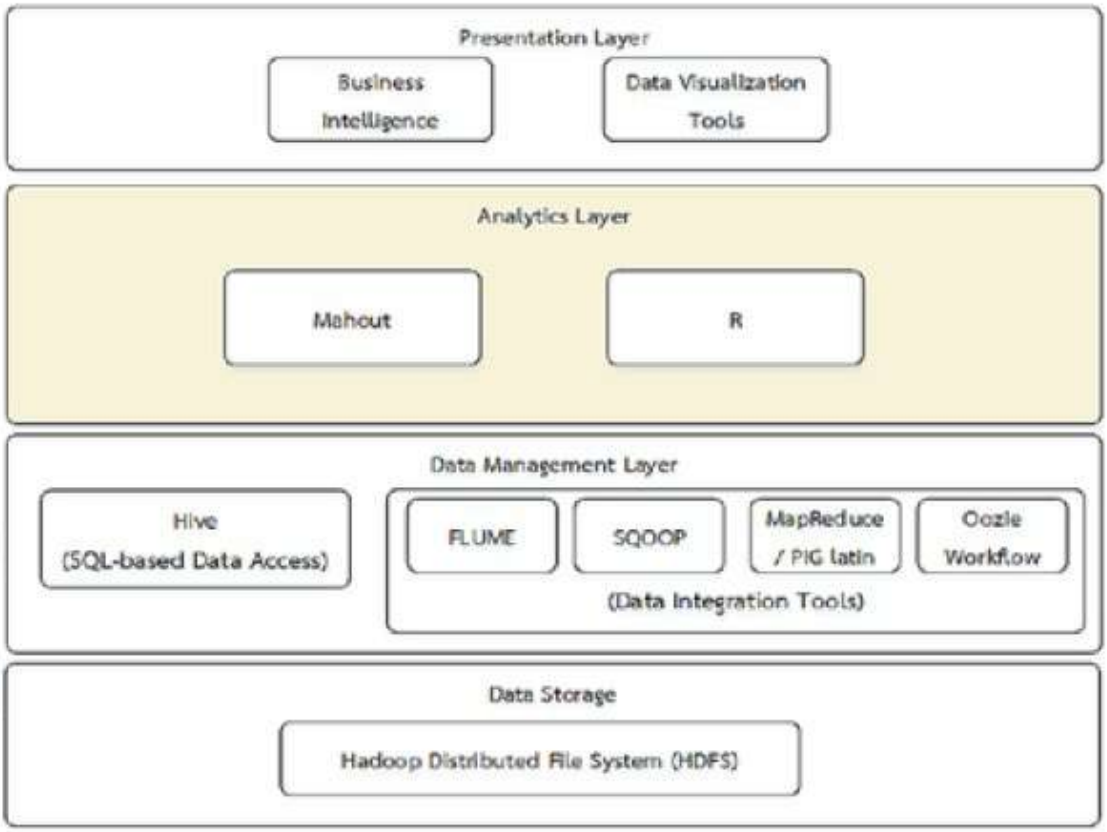



**ภาพที่ 9.1 ตัวอย่างการวิเคราะห์ข้อมูลในการประมวลผลแบบเรียลไทม์**

ที่มา: Eri, Khattak, & Buhler (2016)

*** หมายเหตุ ให้ศึกษาคำอธิบายเพิ่มเติมในเอกสารการสอนของแต่ละรูปแบบการวิเคราะห์ที่ประกอบด้วย 6 รูปแบบ และควรเชื่อมโยงแต่ละรูปแบบการวิเคราะห์กับตัวอย่างที่มีการประยุกต์หรือใช้งานจริงให้ได้

**2.1.2** เครื่องมือที่ใช้สำหรับการวิเคราะห์ข้อมูลขนาดใหญ่ส่วนมากจะนำไปใช้ในกระบวนการตรวจสอบ การทำความสะอาดข้อมูล การแปลง และการสร้างโมเดล โดยมีวัตถุประสงค์เพื่อการค้นหาสารสนเทศจากข้อมูลที่สามารถนำไปใช้ประโยชน์ได้ทั้งในด้านการสนับสนุนการตัดสินใจ การหาข้อสรุป หรือการหาข้อแนะนำ เป็นต้น




**ภาพที่ 9.2 เลเยอร์การวิเคราะห์ ในโครงสร้างของฮาดูป**

*** หมายเหตุ ให้ศึกษาคำอธิบายเพิ่มเติมในเอกสารการสอนของแต่ละเครื่องมือที่ใช้สำหรับการวิเคราะห์ข้อมูลขนาดใหญ่โดยเฉพาะเครื่องมือ R และ Mahout และควรเชื่อมโยงแต่ละเครื่องมือการวิเคราะห์กับตัวอย่างที่มีการประยุกต์หรือใช้งานจริงให้ได้

**2.2** อาร์เป็นซอฟต์แวร์โอเพนซอร์สที่ถูกออกแบบมาเพื่อให้ทำงานบนระบบปฏิบัติการที่หลากหลายสามารถในการประมวลผลคำสั่งทั้งในรูปแบบอนุกรม และแบบขนาน รองรับการวิเคราะห์ข้อมูลขนาดใหญ่ที่ใช้การเรียนรู้ของเครื่อง ทั้งแบบมีผู้สอน และแบบไม่มีผู้สอน อาร์มีแพคเกจไลบรารีสำหรับการวิเคราะห์เพื่อประมวลข้อมูลด้านสถิติและคณิตศาสตร์จำนวนมาก และได้รับการปรับปรุงอย่างต่อเนื่องตลอดเวลา

ในข้อนี้มีประเด็นกล่าวถึงได้แก่

**2.2.1** อาร์เป็นภาษาโปรแกรมอาเรย์ที่เน้นการคำนวณข้อมูลพร้อมกันเป็นกลุ่ม มีความสามารถในการประมวลผลคำสั่งทั้งในรูปแบบอนุกรม และแบบขนาน อาร์เก็บประเภทของข้อมูลเพื่อการใช้งานในรูปแบบที่หลากหลาย ได้แก่ ตัวเลข ตัวอักษร ตรรกะ เป็นต้น โครงสร้างข้อมูลของอาร์มีทั้งแบบที่เราคุ้นเคยกันเป็นอย่างดี ได้แก่ อาเรย์ ลิสต์ รวมทั้งรูปแบบที่เป็นลักษณะพิเศษของอาร์ ได้แก่ เวกเตอร์ เฟรมข้อมูล เป็นต้น อาร์มีฟังก์ชันในตัวที่สร้างไว้ให้ผู้พัฒนาโปรแกรมเรียกใช้งานได้ทันที ฟังก์ชันบางส่วนที่ต้องการใช้งานอาจจะไม่ได้มาพร้อมกับการติดตั้งอาร์ หากต้องการเรียกใช้ต้องทำการติดตั้งแพคเกจที่บรรจุฟังก์ชันเหล่านั้น

**2.2.2** การใช้อาร์วิเคราะห์ข้อมูลด้วยวิธีเรียนรู้แบบมีผู้สอน เป็นหนึ่งในรูปแบบการเรียนรู้ของเครื่องที่สร้างโมเดลจากข้อมูลฝึกหัด เพื่อสร้างตัวแบบที่สามารถนำไปใช้ในการจำแนกประเภทของข้อมูล หรือเป็นการหาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เพื่อสร้างฟังก์ชันที่ให้ค่าต่อเนื่องในลักษณะการถดถอย

การจำแนกประเภทของข้อมูลด้วยการสร้างต้นไม้ตัดสินใจด้วยอาร์ มีลำดับขั้นตอนเบื้องต้นดังนี้ (ตอนที่ 9.2 เรื่อง 9.2.2)

1) ชุดข้อมูลการสร้างต้นไม้ตัดสินใจ (ให้พิจารณามุมมองของชุดข้อมูลประเภทอื่น ถ้าเป็นชุดข้อมูลอื่นต้องมีการเก็บรายการของข้อมูลอย่างไร)
2) ข้อมูลฝึกหัด และข้อมูลทดสอบ (ทำอย่างไร เพราะอะไรต้องทำแบบนี้)
3) แพกเกจที่ใช้สร้างต้นไม้ตัดสินใจ (ทำเพื่ออะไร)
4) การสร้างและการทดสอบโมเดล (ทำเพื่ออะไร แล้วจะได้ผลลัพธ์อย่างไร)
5) ข้อมูลของ Confustion matrix (แสดงผลอย่างไรและแปลความหมายได้อย่างไร)

**2.2.3** การใช้อาร์วิเคราะห์ข้อมูลด้วยวิธีเรียนรู้แบบไม่มีผู้สอน เป็นหนึ่งในรูปแบบการเรียนรู้ของเครื่อง ซึ่งเป็นการสร้างโมเดลที่เหมาะสมกับข้อมูลเพื่อใช้ในการแบ่งกลุ่มของข้อมูล หรือการค้นหาความสัมพันธ์ภายในของข้อมูล เพื่อนำไปหารูปแบบที่เกิดขึ้นบ่อย

การแบ่งกลุ่มของข้อมูล (clustering) ด้วยอาร์ มีลำดับขั้นตอนเบื้องต้นดังนี้ (ตอนที่ 9.2 เรื่องที่ 9.2.3)

1) ชุดข้อมูลการแบ่งกลุ่ม (ให้พิจารณามุมมองของชุดข้อมูลประเภทอื่น ถ้าเป็นชุดข้อมูลอื่นต้องมีการเก็บรายการของข้อมูลอย่างไร)
2) ฟังก์ชันและแพกเกจที่ใช้ในการแบ่งกลุ่ม (มีอะไรบ้าง และทำอย่างไร เพราะอะไรต้องทำแบบนี้)
3) การประมาณจำนวนคลัสเตอร์ที่เหมาะสม (ทำไมต้องทำ และทำอย่างไร)
4) การแบ่งกลุ่ม (ทำหรือกำหนดอย่างไร)
5) การวิชวลไลซ์คลัสเตอร์ (ทำอย่างไร และทำเพื่ออะไร)

2.3 มาฮาวท์เป็นซอฟต์แวร์โอเพนซอร์ส โดยมีเป้าหมายในการสร้างสภาพแวดล้อมอันเหมาะสม ให้ความสะดวกแก่นักพัฒนาโปรแกรมเพื่อนำไปใช้สร้างแอปพลิเคชันที่มีประสิทธิภาพสูง มาฮาวท์เป็นคอมโพเนนต์หนึ่งของฮาดูปที่ใช้ในการวิเคราะห์ข้อมูล สามารถปรับขนาดในการใช้งานตามปริมาณของข้อมูลได้ จึงมักนำไปใช้ประโยชน์โดยเฉพาะในการวิเคราะห์เชิงพยากรณ์กับข้อมูลขนาดใหญ่ และนำไปประยุกต์ใช้ในการสร้างระบบแนะนำ มาฮาวท์รองรับการเรียนรู้ด้วยเครื่องวิธีเรียนรู้ ทั้งแบบแบบมีผู้สอน และแบบไม่มีผู้สอน

ในข้อนี้มีประเด็นกล่าวถึงได้แก่

2.3.1 มาฮาวท์ถูกออกแบบให้รันอยู่บนฮาดูปเฟรมเวิร์ค เพื่อประมวลผลข้อมูลที่เก็บไว้ในเอชดีเอฟเอส หรือทำงานกับข้อมูลที่เก็บอยู่ในหน่วยความจำ มาฮาวท์ถูกสร้างขึ้นมาโดยวัตถุประสงค์หลักเพื่อรองรับการเรียนรู้ด้วยเครื่องในลักษณะที่ปรับขนาดได้ เช่น การจำแนก การจัดกลุ่ม การสร้างกลไกระบบแนะนำ และสามารถในไปใช้ในลักษณะอื่นได้อีกหลากหลาย

2.3.2 มาฮาวท์สามารถนำมาใช้ในวิเคราะห์ข้อมูลด้วยวิธีเรียนรู้แบบมีผู้สอนข้อมูล เช่น การจำแนก หรือการพยากรณ์ได้หลากหลาย ข้อดีของมาฮาวท์ที่เหนือกว่าเครื่องมือตัวอื่นคือความสามารถในการรองรับจำนวนข้อมูลฝึกหัดที่ใช้ในการเทรน โดยเฉพาะที่มีจำนวนของข้อมูลฝึกหัดสูงมาก ๆ การสร้างโมเดลวิธีเรียนรู้แบบมีผู้สอนด้วยมาฮาวท์แบ่งเป็น 3 ระยะ ได้แก่ การเทรนโมเดล การประเมินโมเดล และการนำโมเดลไปใช้งาน

การเรียนรู้แบบมีผู้สอนด้วยการวิเคราะห์ข้อมูลแบบมาฮาวท์ มีเกณฑ์เบื้องต้นในการนำมาฮาวท์มาประยุกต์ใช้ได้แก่ 1) ขนาดของข้อมูลฝึกหัดที่เหมาะสม 2) การสร้างโมเดลวิธีการเรียนรู้ 3) การคัดเลือกตัวแปรทำนายโดยมีอัลกอริธึมสำหรับการจำแนกได้หลายตัวและมีลำดับขั้นตอนเบื้องต้นได้แก่ 1) ชุดข้อมูลที่ใช้ในการจำแนก 2) คำสั่งที่ใช้ในการสร้างโมเดลจำแนก (แนวคิดการกำหนด feature เพื่อสร้างโมเดลทำอย่างไร) 3) การแสดงผลลัพธ์ (ผลลัพธ์เป็นลักษณะอย่างไรและบ่งบอกถึงอะไร)

*** หมายเหตุ ศึกษาเพิ่มเติมในตอนที่ 9.3 เรื่อง 9.3.2

2.3.3 การใช้มาฮาวท์วิเคราะห์ข้อมูลด้วยวิธีเรียนรู้แบบไม่มีผู้สอน สามารถทำได้โดยการแบ่งกลุ่มข้อมูล อัลกอริธึมที่รองรับรูปแบบการแบ่งกลุ่มข้อมูล ได้แก่ เค-มีนส์ ฟัซซีเค-มีนส์ และ แคโนปี เป็นต้น การแบ่งกลุ่มข้อมูลใช้การวัดความเหมือน หรือ ความใกล้ชิดของข้อมูล โดยคำนวณจากระยะระหว่างเวกเตอร์ของข้อมูลเข้า ด้วยวิธีต่าง ๆ ได้แก่ การวัดระยะห่างยูคลิเดียน การวัดระยะห่างยูคลิเดียนยกกำลังสอง การวัดระยะห่างแมนฮัตตัน เป็นต้น

การแบ่งกลุ่มข้อมูลด้วยมาฮาวท์ มีลำดับขั้นตอนเบื้องต้นดังนี้

1) ชุดข้อมูลการแบ่งกลุ่ม (มีลักษณะอย่างไร)
2) ใช้อัลกอริธึมสำหรับการวิเคราะห์ (อัลกอริธึมที่นำมาใช้วิเคราะห์มีอะไรบ้าง แต่ละแบบเป็นอย่างไรและทำการวิเคราะห์ได้อย่างไร)
3) การแสดงผลลัพธ์ (ผลลัพธ์ที่ได้มีลักษณะอย่างไร และอธิบายผลจากการทำงานได้อย่างไร)

*** หมายเหตุ ศึกษาเพิ่มเติมในตอนที่ 9.3 เรื่องที่ 9.3.3 และการหาตัวอย่างอื่นๆ ที่มีการนำไปประยุกต์ได้

**2.3.4** รูปแบบที่นิยมใช้ในการสร้างระบบแนะนำด้วยมาฮาวท์ประกอบด้วย การแนะนำแบบอิงผู้ใช้ และการแนะนำแบบอิงสินค้า กลไกตัวแนะนำจะวิเคราะห์ข้อมูลความพึงใจ เพื่อสร้างคำแนะนำในสิ่งที่คาดว่าผู้ใช้มีความสนใจ หรือเป็นสิ่งที่ผู้ใช้กำลังต้องการ มักนำไปใช้ประโยชน์ในการทำพาณิชย์อิเล็กทรอนิกส์

รูปแบบระบบการให้คำแนะนำมี 2 แบบได้แก่ แบบ user-based recommendation และแบบ iterm-based recommendation โดยในแบบ user-based การสร้างระบบแนะนำด้วยมาฮาว์มีลำดับขั้นตอนเบื้องต้นดังนี้ (ศึกษาเพิ่มเติมในตอนที่ 9.3 เรื่องที่ 9.3.4 และลองหาข้อมูลมาทำการวิเคราะห์)

1) การสร้างข้อมูลสำหรับระบบแนะนำ (มีขั้นตอนทำอย่างไร หรือมีหลักการอย่างไร)
2) แนวคิดในการสร้างระบบแนะนำด้วยมาฮาวท์ (มีขั้นตอนทำอย่างไร)
3) การวิเคราะห์ผลลัพธ์จากการแนะนำ (ผลลัพธ์ที่ได้จากระบบเป็นลักษณะอย่างไร และจะแปลผลลัพธ์หรืออธิบายผลได้อย่างไร)