แนวทางเตรียมสอบ 09600 ประมวลความรู้ วท.ม.ไอซีที ภาค 2/2563

หลักสูตรปรับปรุง พ.ศ. 2560 (นักศึกษารหัส 60 เป็นต้นไป)

ภาคทฤษฎี: (3 ข้อ: ทำทุกข้อ)

99703 การจัดการเทคโนโลยีสารสนเทศและการสื่อสารเชิงกลยุทธ์

- สไลด์ประกอบการบรรยายของอ.ดร.เฉลิมศักดิ์ เลิศวงศ์เสถียร เรื่อง “Information Technology Strategic Planning”
- หน่วยที่ 13 การวางแผนกลยุทธ์เทคโนโลยีสารสนเทศ เรื่องที่ 13.2.1 การวิเคราะห์สภาพแวดล้อม และตารางที่ 13.2 การวิเคราะห์แรงแข่งขันทั้ง 5

99708 ระเบียบวิธีวิจัยและเครื่องมือในการพัฒนาระบบด้านเทคโนโลยีสารสนเทศและการสื่อสาร

- ให้ศึกษาเนื้อหาชุดวิชา 99708 หน่วยที่ 2 และกิจกรรมอีเลินนิ่ง เรื่อง การเขียน use case diagram และ activity diagram
- ให้ทบทวนและฝึกเขียน Use Case Diagram และ Activity Diagram การสร้างแผนภาพคลาส (Class Diagram) การแสดง classes, attributes, operations, associations /aggregations และ generalizations จากกรณีศึกษาที่กำหนดให้ โดยใช้กิจกรรมอีเลินนิงเรื่อง ระบบ Banking application เป็นแนวทางในการศึกษา

99709 ธุรกิจอิเล็กทรอนิกส์และการประยุกต์

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

- เทคโนโลยีแพลตฟอร์ม(Technology Platform) ที่สนับสนุนธุรกิจดิจิทัลได้แก่

1. สื่อสังคม (Social Media) หรือ เทคโนโลยีสื่อสังคม (Social technology) คือสื่ออิเล็กทรอนิกส์ซึ่งเป็นสื่อกลางที่ให้บุคคลทั่วไปมีส่วนร่วมสร้างและแลกเปลี่ยนความคิดเห็นต่างๆ ผ่านอินเทอร์เน็ตได้ เช่น เฟซบุ๊ค (Facebook) ไลน์ เป็นต้น
2. โมบาย (Mobile) คือ อุปกรณ์สื่อสารไร้สายที่เคลื่อนที่ได้ เช่น โทรศัพท์มือถือ หรือสมาร์ทโฟน (Smart Phone) เท็บเล็ต (tablet) หรือ อุปกรณ์ที่มีความสามารถในการติดต่อสื่อสารแบบเคลื่อนที่ได้ (mobility) การใช้โมบายแอพพลิเคชั่น (mobile applications)
3. การวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics) คือ การใช้เครื่องมือวิเคราะห์ข้อมูลจำนวนมหาศาลในเชิงลึก
4. การประมวลผลกลุ่มเมฆ (Cloud Computing) หรือคลาวด์ (Cloud) คือบริการที่ครอบคลุมถึงการให้ใช้กำลังประมวลผล หน่วยจัดเก็บข้อมูล และระบบออนไลน์ต่างๆจากผู้ให้บริการ เพื่อลดความยุ่งยากในการติดตั้ง ดูแลระบบ ช่วยประหยัดเวลา และลดต้นทุนในการสร้างระบบคอมพิวเตอร์และเครือข่ายเอง
5. อินเทอร์เน็ตของสรรพสิ่ง (Internet of Things: IoT) คือ สภาพแวดล้อมอันประกอบด้วยสรรพสิ่งที่สามารถสื่อสารและเชื่อมต่อกันได้ผ่านโพรโทคอลการสื่อสารทั้งแบบใช้สายและไร้สาย โดยสรรพสิ่งต่าง ๆ มีวิธีการระบุตัวตนได้ รับรู้บริบทของสภาพแวดล้อมได้ และมีปฏิสัมพันธ์โต้ตอบและทำงานร่วมกันได้
    - เทคโนโลยีที่ช่วยให้สรรพสิ่งรับรู้ข้อมูลในบริบทที่เกี่ยวข้อง เช่น เซ็นเซอร์ ระบบสมองกลฝังตัว

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

ธุรกิจดิจิทัลทำให้ธุรกิจเดิมค่อย ๆ เสื่อมความนิยมเนื่องจากผู้บริโภคได้รับความสะดวกและคุณค่าจากธุรกิจดิจิทัลมากกว่า ธุรกิจจำเป็นต้องรีบเร่งปรับตัวและเตรียมความพร้อมที่จะรับมือกับอิทธิพลของเทคโนโลยีดิจิทัลและการแข่งขันรูปแบบใหม่ การปรับเปลี่ยนเพื่อให้เป็นธุรกิจดิจิทัลที่สมบูรณ์แบบมีขั้นตอนครอบคลุมเรื่องต่าง ๆ อย่างน้อย 5 เรื่องคือ 1) ทำธุรกรรมด้วยเอกสารอิเล็กทรอนิกส์ 2) เพิ่มช่องทางการทำธุรกรรมผ่านสื่ออิเล็กทรอนิกส์ 3) ส่งเสริมการสร้างนวัตกรรมบริการ (Service Innovation) 4) ใช้เทคโนโลยีดิจิทัลเพื่อขยายธุรกิจสู่ตลาดโลก และ 5) สร้างประสบการณ์ที่ดีให้ลูกค้า (Customer Experience) **ขอให้นักศึกษาไปศึกษารายละเอียดที่เกี่ยวข้องในส่วนนี้**

ภาคประยุกต์: (3 ข้อ: บังคับทำข้อสอบ 1) 99710 2) ข้อสอบบูรณการ และ 3) เลือกทำ 99705 หรือ 99707 หรือ 99711)

99710 เทคโนโลยีเคลื่อนที่ไร้สายและการประยุกต์ **(ข้อบังคับ)**

- แนวทางการประยุกต์ใช้เทคโนโลยีเคลื่อนที่ไร้สายและเทคโนโลยีสื่อสารอื่นๆ ที่เกี่ยวข้องสำหรับเมืองอัจฉริยะด้านต่างๆ (Smart City) ผ่านคลาวด์คอมพิงทิง บิ๊กดาตา และอุปกรณ์อิเล็กทรอนิกส์เคลื่อนที่ไร้สายต่างๆ

เลือกทำ

99705 ความมั่นคงด้านเทคโนโลยีสารสนเทศและการสื่อสาร

- ศึกษาประเภทและรูปแบบการโจมตีระบบเครือข่ายคอมพิวเตอร์ลักษณะต่างๆ ได้ พร้อมยกตัวอย่างและเสนอแนะแนวทางการป้องกัน

99707 ระบบสารสนเทศภูมิศาสตร์และการประยุกต์

ภูมิสารสนเทศเป็นศาสตร์ที่ครอบคลุมถึงระบบสารสนเทศภูมิศาสตร์ ภูมิสารสนเทศ จึงเป็นวิทยาศาสตร์และเทคโนโลยีที่เกี่ยวข้องกับการได้มา (Capture) การบูรณาการ (Integrating) การวิเคราะห์ (Analyzing) การจัดการ (Managing) และ การตีความ (Depicting) ข้อมูลข่าวสารเชิงพื้นที่ อันประกอบไปด้วยข้อมูล 3 ด้าน คือ 1) ทำเลที่ตั้ง (Location) ที่สามารถบอกเป็นค่าพิกัดที่แน่นอนได้ 2) สภาพแวดล้อมทางธรรมชาติเป็นข้อมูลที่แสดงถึงสิ่งแวดล้อมที่เกิดขึ้นเองตามธรรมชาติและ 3) สภาพแวดล้อมทางวัฒนธรรม เป็นข้อมูลที่แสดงถึงสิ่งแวดล้อมที่มนุษย์สร้างขึ้น

ประเภทข้อมูลในระบบสารสนเทศภูมิศาสตร์ประกอบด้วยข้อมูล 2 ประเภท คือ 1)ข้อมูลเชิงพื้นที่ (Spatial data)เป็นข้อมูลที่เกี่ยวข้องกับตำแหน่งที่ตั้งของข้อมูลต่าง ๆ บนพื้นโลก สามารถอ้างอิงกับตำแหน่งทางภูมิศาสตร์แสดงสัญลักษณ์ได้ 3 รูปแบบ คือ จุด เส้น และพื้นที่ 2) ข้อมูลเชิงคุณลักษณะ (Non-ซึ่งเป็นลักษณะข้อมูลเชิงบรรยาย โดยจะอธิบายถึงคุณลักษณะต่าง ๆ ในพื้นที่นั้น ให้ศึกษารายละเอียดเพิ่มเติม

ความแตกต่างระหว่างข้อมูลของระบบสารสนเทศทั่วไปกับข้อมูลของระบบสารสนเทศภูมิศาสตร์ รวมถึงการใช้ประโยชน์เป็นสิ่งที่นักศึกษาต้องไปศึกษารายละเอียดเพิ่มเติม

ฐานข้อมูลระบบสารสนเทศภูมิศาสตร์มีลักษณะการเชื่อมโยงระหว่างฐานข้อมูลเชิงบรรยายกับข้อมูลเชิงเส้นหรือข้อมูลเชิงภาพ กระบวนการในการวิเคราะห์ข้อมูลเชิงพื้นที่ มีกระบวนการหลัก คือ การกำหนดจุดประสงค์ การวางแผนวิเคราะห์รวบรวมข้อมูล ทำการวิเคราะห์ และสรุปผล

กระบวนการในการวิเคราะห์ ประกอบด้วย **1.กำหนดจุดประสงค์** หรือระบุปัญหาที่ต้องการหาคำตอบให้ชัดเจนจัดลำดับความสำคัญของจุดมุ่งหมายให้ชัดเจน (set priority) ว่าอะไรคือจุดประสงค์หลัอะไรคือจุดประสงค์รอง สิ่งใดมีความสำคัญหรือสิ่งใดที่เป็นเพียงผลพลอยได้จากการวิเคราะห์ **2.วางแผนการวิเคราะห์และรวบรวมข้อมูล** เป็นการศึกษาและรวบรวมปัจจัยในการวิเคราะห์รวมไปถึงสถิติลักษณะ คุณภาพและข้อจำกัดของข้อมูลที่มี การจัดกลุ่มข้อมูล เลือก modelที่เหมาะสมในการวิเคราะห์ประมาณงบประมาณที่จะใช้ ระบุความต้องการ ความเสี่ยง หรือรายละเอียดของข้อมูลที่จะช่วยในการตัดสินใจ และระยะเวลาการวิเคราะห์ ทรัพยากรที่ต้องใช้ ปัจจัยเสี่ยงและความเป็นไปได้ มาตราฐานแหล่งที่มาของข้อมูล การสุ่มตัวอย่าง ข้อจำกัดและการทบทวนปัญหาที่ต้องการคำตอบในชั้นต้น **3. ทำการวิเคราะห์** ในการวิเคราะห์ข้อมูลนั้นอาจมีความจำเป็นต้องแบ่งการวิเคราะห์ออกเป็นประเด็นย่อยๆ ตามจุดประสงค์ ขั้นตอน หรือmodel ที่ได้วางไว้ในตอนแรกเพื่อให้ง่ายในการวิเคราะห์ และเพื่อให้การ

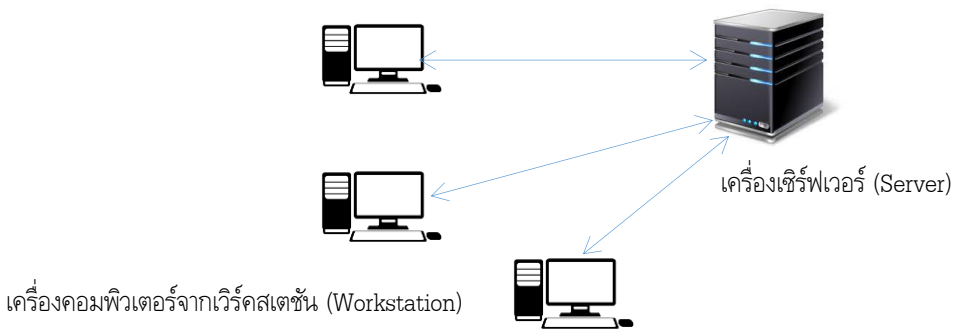
วิเคราะห์นั้นมีความถูกต้องและชัดเจนมากที่สุด ผู้วิเคราะห์สามารถเลือกใช้โปรแกรมประยุกต์แบบสำเร็จรูปที่มี module หรือ function พร้อมใช้เพื่อความสะดวก หรือโปรแกรมแบบที่สามารถเขียน module หรือ function เพิ่มเติมเอง ได้ตามความต้องการ ไม่ว่าจะเป็น Arc GIS, Quantum GIS, GRASS GIS หรือโปรแกรม R4 **สรุปผล** กระบวนการสุดท้ายของกระบวนการวิเคราะห์ข้อมูลก็คือการสรุปผล ไม่ว่าจุดประสงค์หรือสมมุติฐานที่ที่ตั้งในเบื้องต้นนั้นจะได้รับคำตอบหรือไม่ อย่างไร เป็นไปตามความคาดหมายในเบื้องต้นหรือไม่ มีข้อดี ข้อด้อย ข้อจำกัดหรือขอบเขต คุณภาพ ความถูกต้อง และการประเมินขั้นตอนในการวิเคราะห์อย่างไร ผลลัพธ์ที่ได้ยอมรับได้หรือไม่ คำแนะนำต่อไปเป็นอย่างไร ต้องทำการวิเคราะห์ต่อหรือไม่ **ให้นักศึกษาศึกษารายละเอียดเพิ่มเติมในฟังก์ชั่นของการวิเคราะห์ข้อมูล**

99711 การวิเคราะห์ข้อมูลขนาดใหญ่สำหรับธุรกิจ

แมปรีดิวซ์เป็นกรอบการทำงานที่ทำงานอยู่บนฮาดูป ที่นำข้อมูลที่ได้แบ่งแยกให้เป็นข้อมูลเล็ก เข้าสู่ขั้นตอนการทำงานของแมปรีดิวซ์ โดยการดำเนินการของแมปรีดิวซ์ประกอบด้วย 6 ขั้นตอนหลัก ได้แก่ ข้อมูลนำเข้า การแบ่งแยกข้อมูล การแมป การสับเปลี่ยน/จัดเรียง การรีดิวซ์ และข้อมูลผลลัพธ์ โดยประโยชน์ของแมปรีดิวซ์ ได้แก่ ความสามารถในการเพิ่มขยายได้ในอนาคต ประหยัดค่าใช้จ่าย มีความยืดหยุ่นสูง การประมวลผลข้อมูลรวดเร็วและความทนทานต่อความผิดพลาด

1 ความเป็นมาและความหมายของแมปรีดิวซ์

1.1 ความเป็นมาของแมปรีดิวซ์ จากระบบการบริหารจัดการข้อมูลแบบดั้งเดิมขององค์กร โดยปกติจะมีเครื่องเซิร์ฟเวอร์ส่วนกลางในการจัดเก็บและประมวลผลข้อมูล ดังภาพที่ 6.1



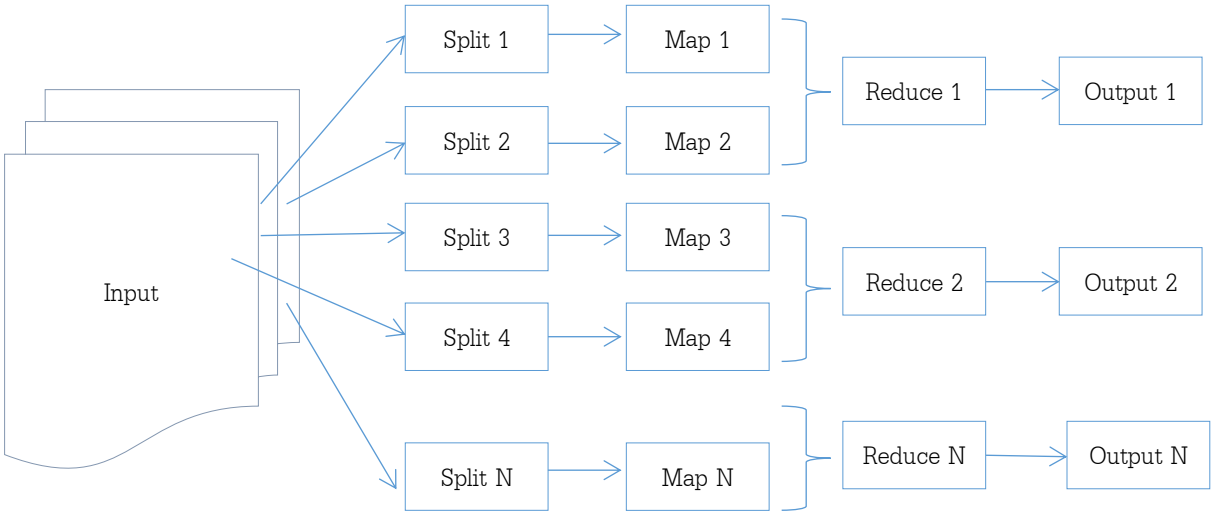
ภาพที่ 6.1 ระบบการบริหารจัดการข้อมูลแบบบระบบรวมศูนย์ (Centralized System)

ภาพที่ 6.1 ระบบการบริหารจัดการข้อมูลแบบบระบบรวมศูนย์ (Centralized System) เป็นวิธีการบริหารจัดการข้อมูลโดยผู้ใช้เครื่องคอมพิวเตอร์ทั้งจากเวิร์คสเตชัน (Workstation) หรือเครื่องปลายทาง (terminal) ต่างๆ ทำการเชื่อมต่อผ่านระบบเครือข่ายไปที่เครื่องเซิร์ฟเวอร์ (Server) ที่ศูนย์กลางเพียงจุดเดียว ในการจัดการข้อมูล อาทิการเพิ่ม

รายการข้อมูลการเปลี่ยนแปลงต่างๆ ซึ่งเป็นระบบการบริหารจัดการข้อมูลแบบดั้งเดิมขององค์กรต่างๆ ซึ่งอาจจะไม่เหมาะในการประมวลผลข้อมูลที่มีขนาดใหญ่และจำนวนมาก (Large Volume of Data) เช่นข้อมูลจาก Facebook หรือ Youtube และข้อมูลเกิดการเปลี่ยนแปลงตลอดเวลา จึงไม่สามารถเก็บลงในเครื่องเซิร์ฟเวอร์ส่วนกลางและใช้ระบบการจัดการฐานข้อมูลเชิงสัมพันธ์ (Relational Database Management System) ได้ เนื่องจากข้อมูลมีความหลากหลาย นอกจากในการจัดการข้อมูลแบบระบบรวมศูนย์ที่ทำการเชื่อมต่อเครื่องคอมพิวเตอร์ทั้งจากเวิร์คสเตชันและเครื่องปลายทางหลายๆ เครื่องให้มารวมศูนย์ที่เครื่องเซิร์ฟเวอร์จุดเดียวจะก่อให้เกิดปัญหาคอขวดในขณะที่ประมวลผลไฟล์ข้อมูลจากเครื่องคอมพิวเตอร์ในระบบเครือข่ายหลายเครื่องพร้อมกัน

จากโจทย์ปัญหาวิธีการบริหารจัดการข้อมูลแบบบระบบรวมศูนย์ ซึ่งก่อให้เกิดปัญหาคอขวด ทางบริษัทกูเกิล (Google) ได้คิดแก้ไขปัญหาคอขวดนี้โดยใช้ขั้นตอนวิธีการที่เรียกว่า แมปรีดิวซ์ (MapReduce) โดยแบ่งแยกข้อมูลออกเป็นส่วนเล็ก ๆ และกำหนดให้เครื่องคอมพิวเตอร์จำนวนมาก นำข้อมูลส่วนเล็ก ๆ ไปประมวลผลและบูรณาการผลลัพธ์กลับคืนมา เพื่อเพิ่มประสิทธิภาพในการประมวลผลข้อมูลใหญ่ (Big Data) ทำให้การประมวลผลรวดเร็วมากยิ่งขึ้น และใช้ทรัพยากรเครื่องคอมพิวเตอร์ให้คุ้มค่ามากที่สุด โดยไม่จำเป็นต้องจัดหาเครื่องเซิร์ฟเวอร์ที่มีสมรรถนะสูงหรือราคาแพงจำนวนมากอาจจะเป็นเครื่องคอมพิวเตอร์ธรรมดาหลายๆ เครื่อง (Commodity Hardware) ก็สามารถช่วยกันประมวลผลได้ ช่วยให้ข้อมูลที่จัดเก็บมีความปลอดภัยสูงสุด มีความทนทานสูงสุด สามารถซ่อมแซมได้ ทำให้การเก็บข้อมูลขนาดใหญ่หรือการจัดเก็บข้อมูลแบบไม่มีโครงสร้างเป็นไปได้ภายใต้งบประมาณที่จำกัด

1.2 ความหมายของแมปรีดิวซ์ แมปรีดิวซ์คือกรอบการทำงานที่ทำงานอยู่บนฮาดูป โดยแบ่งขั้นตอนการทำงานออกเป็น 2 ส่วนได้แก่ 1) ขั้นตอนแมป (Map Phase) และ 2) ขั้นตอนรีดิวซ์ (Reduce Phase) โดยหลักการทำงานของแมปรีดิวซ์ เริ่มต้นจากที่ขั้นตอนแมป จะทำการแบ่งแยกข้อมูล (Data Splitting) ขนาดใหญ่ให้เป็นชิ้นเล็กๆ แล้วนำข้อมูลชิ้นเล็กๆ หรือที่เรียกว่าบล็อก (Block) กระจายไปประมวลผลบนเครื่องคอมพิวเตอร์แบบคลัสเตอร์ (Computer Clusters) ที่เชื่อมโยงและถูกควบคุมจากคอมพิวเตอร์หลัก (Master Computer หรือ Master Node) หลังจากนั้นผลลัพธ์ที่ได้ของขั้นตอนแมปจะถูกส่งเข้ามาที่ขั้นตอนรีดิวซ์ เพื่อทำการรวบรวมผลการดำเนินงานจากคอมพิวเตอร์เครื่องต่างๆ เข้าเป็นชิ้นเดียวกันและแสดงผลลัพธ์ (Alex, 2012), (Tom, 2012) ดังภาพที่ 6.2

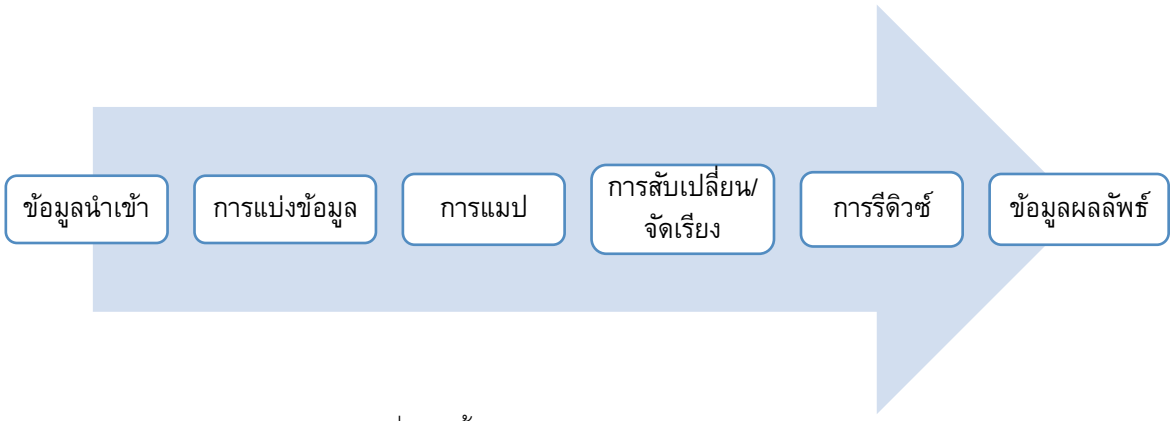


ภาพที่ 6.2 ภาพรวมการดำเนินงานของแมปรีดิวซ์

จากภาพที่ 6.2 ขั้นตอนการดำเนินการของแมปรีดิวซ์ โดยการนำเข้าไฟล์ข้อมูลขนาดใหญ่ที่มีจำนวนมาก มาทำการแบ่งแยก (Split) ให้เป็นข้อมูลเล็ก ได้แก่ Split 1, 2, ..., N ตามลำดับ แล้วนำข้อมูลเล็กๆ เหล่านั้น เข้าสู่ขั้นตอนแมปรีดิวซ์ ซึ่งประกอบด้วย 2 ส่วนได้แก่ 1) ขั้นตอนแมป เป็นการกำหนดคู่คีย์ (Key Pair) หรือคีย์ที่เป็นคู่ของข้อมูลเพื่อทำการกระจายข้อมูลไปประมวลผลยังโหนด (Nodes) ต่างๆ ตามคำสั่งการทำงาน ซึ่งจากภาพ 6.2 เมื่อทำการแบ่งแยกข้อมูล จะนำข้อมูลไปทำการกำหนดคีย์ของค่าข้อมูลในแต่ละแมป อาทิ Map1, 2,..., N และ 2) ขั้นตอนรีดิวซ์ เป็นการนำผลลัพธ์ที่ได้จากการประมวลผลของขั้นตอนแมปจากการทำงานของโหนดต่างๆ (Output ของขั้นตอนแมป มาเป็น Input ของขั้นตอนรีดิวซ์) มาทำการจัดเรียงลำดับและสรุปผลข้อมูล จากภาพที่ 6.2 จึงมีจำนวนรีดิวซ์ ตั้งแต่ Reduce 1, 2, ..., N เพื่อแสดงผลลัพธ์การทำงานที่รวดเร็วในการประมวลผลข้อมูลขนาดใหญ่ ประโยชน์ที่สำคัญของการนำแมปรีดิวซ์ไปประยุกต์ใช้ในกรณีที่สามารถปรับขนาดการประมวลผลบนเครื่องคอมพิวเตอร์แบบคลัสเตอร์ (Computer Clusters) ที่อาจจะประกอบด้วยโหนดตั้งแต่หลักร้อยถึงหมื่นเครื่องได้

2 การดำเนินงานของแมปรีดิวซ์

2.1 ขั้นตอนการดำเนินงานของแมปรีดิวซ์ ในการดำเนินงานของแมปรีดิวซ์ตั้งแต่เริ่มต้นจนสิ้นสุด ประกอบด้วย 6 ขั้นตอนหลัก ดังภาพที่ 6.3

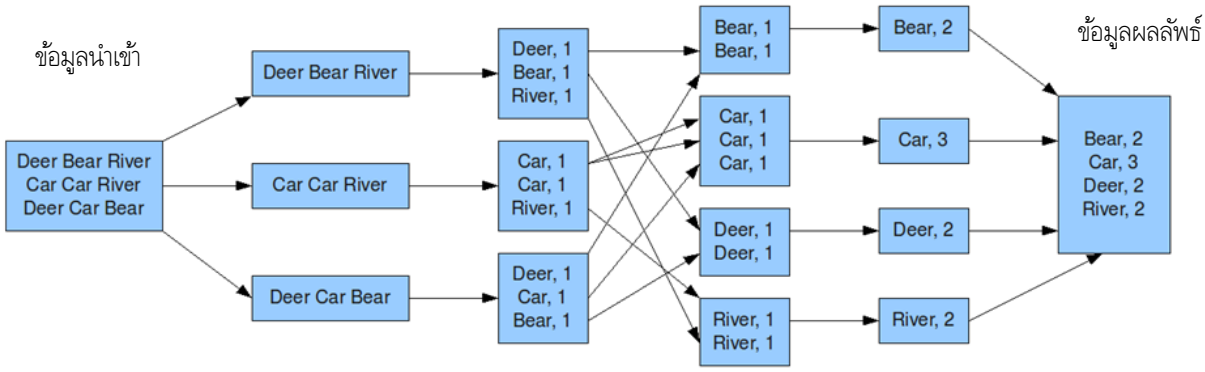


ภาพที่ 6.3 ขั้นตอนการดำเนินงานของแมปรีดิวซ์

จากภาพที่ 6.3 ขั้นตอนการดำเนินงานของแมปรีดิวซ์ ประกอบด้วย 6 ส่วนหลัก เริ่มตั้งแต่ 1) ข้อมูลนำเข้า (Input) 2) การแบ่งข้อมูล (Splitting) 3) การแมป (Mapping) 4) การสับเปลี่ยน/จัดเรียง (Shuffling/Sorting) 5) การรีดิวซ์ (Reducing) และ 6) ข้อมูลผลลัพธ์ (Output)

ด้วยแมปรีดิวซ์ เป็นที่นิยมแพร่หลายในการเพิ่มประสิทธิภาพการประมวลผลข้อมูลขนาดใหญ่บนฮาดูป จึงมีการนำแมปรีดิวซ์ ไปประยุกต์การทำงาน อาทิ โปรแกรมการนับคำ ดังภาพที่ 6.4

การแบ่งข้อมูล การแมป การสับเปลี่ยน/จัดเรียง การรีดิวซ์



ภาพที่ 6.4 ตัวอย่างการดำเนินงานของแมปรีดิวซ์กับโปรแกรมการนับคำ (WordCount) (Sandeep Karanth, 2014)

จากภาพที่ 6.4 ตัวอย่างการดำเนินงานของแมปรีดิวซ์กับโปรแกรมการนับคำ (WordCount) ที่ทำการนับจำนวนคำในเอกสารขนาดใหญ่ โดยทำการตัดแบ่งข้อมูลของเอกสารให้เป็นข้อมูลเล็กๆ และกระจายเอกสารเล็กๆ เหล่านี้ไปประมวลผลตามขั้นตอนของแมปรีดิวซ์ เพื่อให้แต่ละโหนดทำการนับคำเฉพาะในเอกสารข้อมูลเล็กๆ จากนั้นขั้นตอน รีดิวซ์จะนำผลลัพธ์ของจำนวนคำมาทำการจัดเรียงและสรุปผลการทำงาน ว่าในเอกสารนี้มีคำว่าอะไรบ้าง จำนวนกี่คำ ซึ่งจะทำให้ได้ผลลัพธ์ที่รวดเร็วอย่างมาก รายละเอียดดังต่อไปนี้

1) ข้อมูลนำเข้า เป็นไฟล์ข้อมูลเข้ามี 1 ไฟล์และประกอบด้วย 3 ประโยค เช่นประโยคแรก “Deer Bear River” ประโยคที่สอง “Car Car River” และประโยคที่สาม “Deer Car Bear” เป็นต้น

2) การแบ่งข้อมูล เป็นการแบ่งไฟล์ข้อมูลให้มีขนาดเล็กเป็นบรรทัด เช่นข้อมูลที่เป็นไฟล์ Text จะทำการแบ่งข้อมูลจากการขึ้นบรรทัดใหม่ หรือการกด Enter ซึ่งจะแบ่งเป็น 3 ส่วนได้แก่ส่วนที่ 1 : Dear Bear River ส่วนที่ 2 : Car Car River และ ส่วนที่ 3 : Dear Car Bear

3) การแมป เป็นการนับความถี่ของคำจากข้อมูลในแต่ละประโยค เช่นประโยคแรก “Deer Bear River” จะมีค่าคีย์คือ "Deer" และมีค่าของข้อมูลคือ 1 หรือค่าคีย์คือ "Bear" และมีค่าของข้อมูลคือ 1 เป็นต้น

4) การสับเปลี่ยน/จัดเรียง เป็นการนำข้อมูลออก (Output) มาทำการรวบรวมข้อมูลและจัดเรียงลำดับข้อมูลผ่าน http ตามคู่ค่าคีย์ เช่นการสับเปลี่ยน/จัดเรียง ตามตัวอักษรภาษาอังกฤษ A-Z โดย Bear ถูกจัดเรียงให้อยู่ในลำดับแรกและแสดงข้อมูลการนับคำว่า "Bear" ในเอกสารทั้งหมดจะได้ Bear, 1 และ Bear, 1

5) การรีดิวซ์ เป็นการรวบรวมผลลัพธ์จากคำสั่งการทำงานของฟังก์ชันแมป มาแสดงผลลัพธ์ เช่นการรีดิวซ์โดยการหาผลสรุปรวมจากการนับจำนวนคำ เช่นคำว่า "Bear" มี 2 คำ และคำว่า "Car" มี 3 คำเป็นต้น

ข้อมูลผลลัพธ์ เป็นไฟล์ข้อมูลผลลัพธ์การทำงาน การนับจำนวนคำ โดยมีคำว่า "Bear", 2 คำ คำว่า "Car", 3 คำ คำว่า "Deer", 2 คำ และคำว่า "River", 2 คำ โดยผลลัพธ์จะถูกจัดเรียงลำดับตามตัวอักษรจากน้อยไปมาก

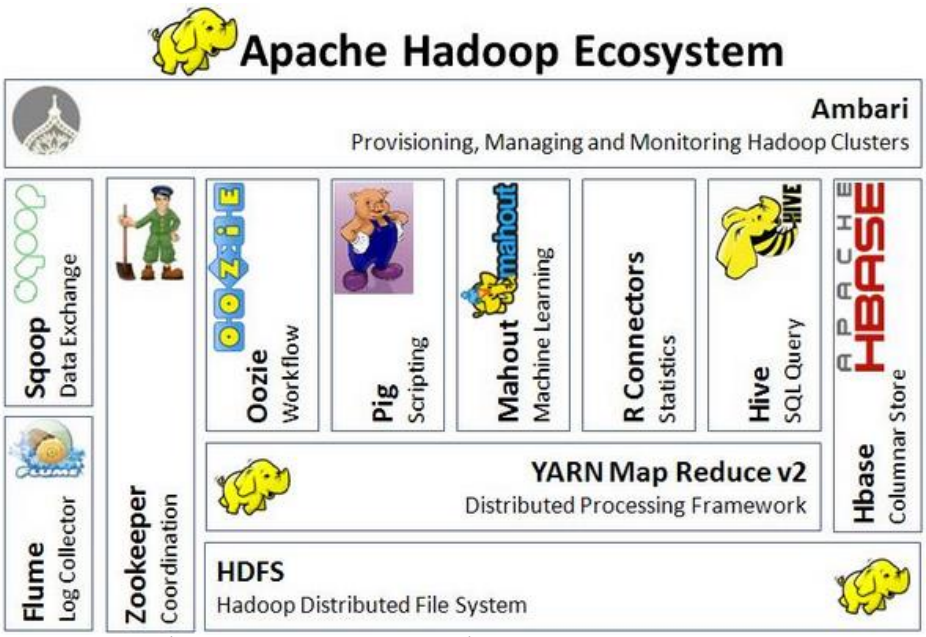
ระบบนิเวศฮาดูปเป็นการรวบรวมซอฟต์แวร์เฉพาะด้านต่างๆ ที่ทำงานอยู่บนอาปาเช ฮาดูป ซึ่งซอฟต์แวร์เหล่านี้เป็นแบบซอฟต์แวร์โอเพนซอร์สของมูลนิธิอาปาเชซอฟต์แวร์ และปัจจุบันมีเทคโนโลยีหลักภายในระบบนิเวศที่เกี่ยวข้องกับฮาดูป ซึ่งมีซอฟต์แวร์หลายประเภทและหน้าที่การทำงานที่แตกต่างกัน เพื่อช่วยอำนวยความสะดวกในการจัดการข้อมูลขนาดใหญ่

ในแต่ละรูปแบบของงาน ได้แก่ฟลูม สคูป ซูคีปเปอร์ อูซี พิก มาฮาวท์ อาร์ ไฮฟ เอชเบส ยาน แฟ้มข้อมูลแบบกระจายฮาดูปและแอมบารี

1 ระบบนิเวศของฮาดูป

ระบบนิเวศฮาดูป (Hadoop ecosystem) เป็นการรวบรวมซอฟต์แวร์เฉพาะด้านต่างๆ ที่ทำงานอยู่บนอาปาเช ฮาดูป ซึ่งซอฟต์แวร์เหล่านี้เป็นแบบซอฟต์แวร์โอเพนซอร์สของมูลนิธิอาปาเชซอฟต์แวร์ โดยปกติการทำงานบนอาปาเชฮาดูปประกอบด้วยหลักการทำงาน 2 ส่วนได้แก่ระบบแฟ้มข้อมูลแบบกระจายฮาดูปและแมปรีดิวซ์ ซึ่งกระบวนการทำงานบนอาปาเชฮาดูปตั้งแต่การกำหนดข้อมูลนำเข้า การประมวลผล หรือการแสดงผลลัพธ์ออกมา จะต้องมีการกำหนดค่าข้อมูลการทำงาน มีการเขียนคำสั่งและมีการปรับแต่งค่าข้อมูลต่างๆ ซึ่งเป็นเรื่องที่ยุ่งยาก ซับซ้อนและอาจจะไม่สะดวกสำหรับผู้ใช้งานทั่วไป ดังนั้นระบบนิเวศฮาดูปจะเป็นการนำเสนอเทคโนโลยีและซอฟต์แวร์ต่างๆ ที่มีหน้าที่การทำงานเฉพาะด้านที่แตกต่างกันออกไป มาประยุกต์ใช้งานเพื่ออำนวยความสะดวกแก่ผู้ใช้งานและเพิ่มประสิทธิภาพในการทำงานให้ดียิ่งขึ้น เช่น มีซอฟต์แวร์ช่วยในการนำข้อมูลนำเข้าสู่ระบบได้สะดวกหรือแปลงรูปแบบข้อมูลให้สะดวกในการประมวลผล การเขียนคำสั่งประมวลผลโดยใช้ภาษาเอสคิวแอล (SQL) การเขียนคำสั่งหรืออ่านข้อมูลแบบสุ่มเพื่อเข้าถึงข้อมูลได้รวดเร็วมากยิ่งขึ้น รวมถึงการถ่ายโอนข้อมูลมาทำงานบนฮาดูป

โดยซอฟต์แวร์ในระบบนิเวศของฮาดูป ที่จะมาช่วยเสริมในเรื่องการจัดการข้อมูล การเข้าถึงและดึงข้อมูล รวมทั้งการติดต่อแลกเปลี่ยนข้อมูลกับระบบต่างๆ ให้สะดวกขึ้นประกอบด้วยซอฟต์แวร์หลัก ได้แก่ซูคีปเปอร์ ไฮฟ พิก สคูป เอชเบส และมาฮาวท์ ดังภาพที่ 7.4



ภาพที่ 7.4 ระบบนิเวศของฮาดูป (ที่มา The Hadoop Ecosystem : https://www.cloudera.com/more/training/library/hadoop-ecosystem.html)

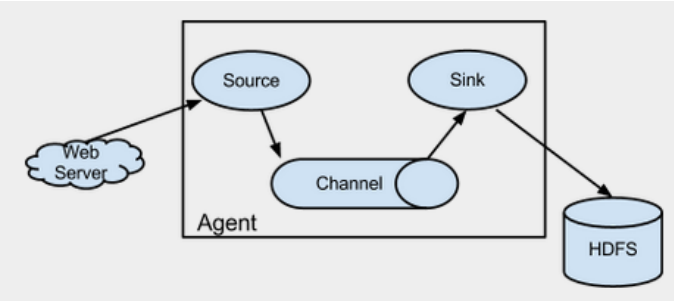
จากภาพที่ 7.4 ระบบนิเวศที่เกี่ยวข้องกับฮาดูปเริ่มต้นที่การนำข้อมูลขนาดใหญ่เข้ามาสู่ขั้นตอนการประมวลผลการทำงานบนอาปาเชฮาดูป ที่มีการจัดการข้อมูลเป็นแบบระบบแฟ้มข้อมูลแบบกระจายฮาดูป และการเขียนคำสั่งการทำงานแบบ

แมปรีดิวซ์ จากนั้นในส่วนการจัดการข้อมูลขนาดใหญ่ประกอบด้วยซอฟต์แวร์หลัก ได้แก่ ซูคีปเปอร์ ทำหน้าที่จัดการประสานการทำงานระหว่างซอฟต์แวร์ต่างๆ ภายในระบบนิเวศที่เกี่ยวข้องของฮาดูป ไฮฟทำหน้าที่จัดการเรื่องคลังข้อมูล พิกทำหน้าที่จัดการการเขียนคำสั่งแบบภาษาสคริปต์ (Script) สคูปทำหน้าที่จัดการเรื่องการถ่ายโอนข้อมูลที่มีโครงสร้าง (Structured Data) เอชเบสทำหน้าที่จัดการเรื่องการอ่านและเขียนข้อมูลแบบเวลาจริง (Realtime) และมาฮาวท์ทำหน้าที่จัดการเรื่องการวิเคราะห์ข้อมูลเชิงทำนายผล โดยซอฟต์แวร์ต่างๆ เหล่านี้จะทำหน้าที่เฉพาะด้านหรือรองรับการทำงานแต่ละเรื่องที่แตกต่างกัน เพื่อให้ตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น และผลลัพธ์ที่ได้จากการจัดการข้อมูลขนาดใหญ่ด้วยเทคโนโลยีฮาดูปหรือซอฟต์แวร์ของระบบนิเวศฮาดูป จะนำไปประยุกต์ใช้ในการทำธุรกิจอัจฉริยะ เช่นการจัดทำรายงานธุรกิจในรูปแบบต่างๆ ที่เหมาะสมกับมุมมองในการวิเคราะห์ แสดงความสัมพันธ์ การทำนายผลลัพธ์ของแนวโน้มที่อาจเกิดขึ้นได้ตรงตามความต้องการขององค์กร เพื่อประโยชน์ในการวางแผนกลยุทธด้านต่างๆ การพยากรณ์ยอดการขายสินค้าหรือการพยากรณ์แนวโน้มทางการตลาด เป็นต้น

2 เทคโนโลยีระบบนิเวศของฮาดูป

รายละเอียดกลุ่มของเทคโนโลยีหลักภายในระบบนิเวศที่เกี่ยวข้องกับฮาดูป ประกอบด้วย 12 ซอฟต์แวร์ดังต่อไปนี้

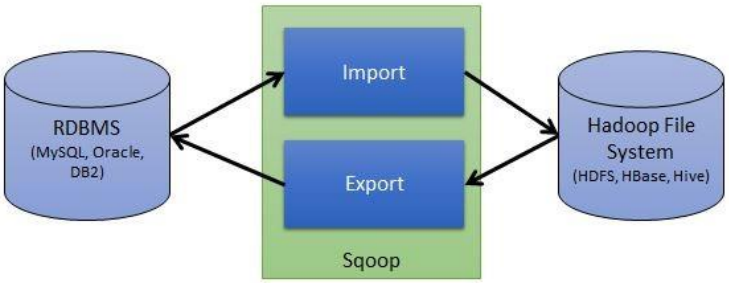
1.1 ฟลูม (Flume) มีสัญลักษณ์ของซอฟต์แวร์เป็นรูปท่อนซุง (Log) โดยซอฟต์แวร์ฟลูมทำหน้าที่จัดการเรื่องเครื่องมือที่ช่วยจัดเก็บข้อมูล Log collector หรือ Event log จากแหล่งต่างๆ โดยทำการเคลื่อนย้ายข้อมูลที่มีจำนวนมากเพื่อไปเก็บไว้ที่ HDFS ดังภาพที่ 7.5



ภาพที่ 7.5 หลักการทำงานของฟลูม (ที่มา: https://flume.apache.org)

จากภาพที่ 7.5 ฟลูมนำข้อมูลเข้าไปยัง Web server ที่มีการฝัง Agent เพื่อทำการไหลข้อมูลผ่านทาง Channel ซึ่งทำงานบน Memory ไปเก็บไว้บน HDFS ทำให้มีความรวดเร็วและทำงานแบบ Real-time ได้ เช่นการประมวลผลข้อมูลของ Web log หรือ Twitter เป็นต้น

1.2 สคูป (Sqoop) มีสัญลักษณ์ของซอฟต์แวร์เป็นตัวอักษรคำว่า Sqoop สีเขียว โดยซอฟต์แวร์สคูปทำหน้าที่จัดการเรื่องการถ่ายโอนข้อมูลที่มีโครงสร้าง (Structured data) เช่นข้อมูลจากระบบฐานข้อมูลที่อยู่รูปแบบตาราง แบบฐานข้อมูลเชิงสัมพันธ์ หรือ Relational Database Management System : RDBMS ได้แก่ SQL server, Oracle หรือ MySQL เข้ามาเก็บในรูปแบบระบบแฟ้มข้อมูลแบบกระจายฮาดูป (Hadoop Distributed File System: HDFS) ซึ่งจะต้องสร้าง Connection ผ่าน JDBC เพื่อเชื่อมต่อไปยังระบบฐานข้อมูลและสร้าง Link เพื่อเชื่อมต่อไปยัง HDFS) ดังภาพที่ 7.6

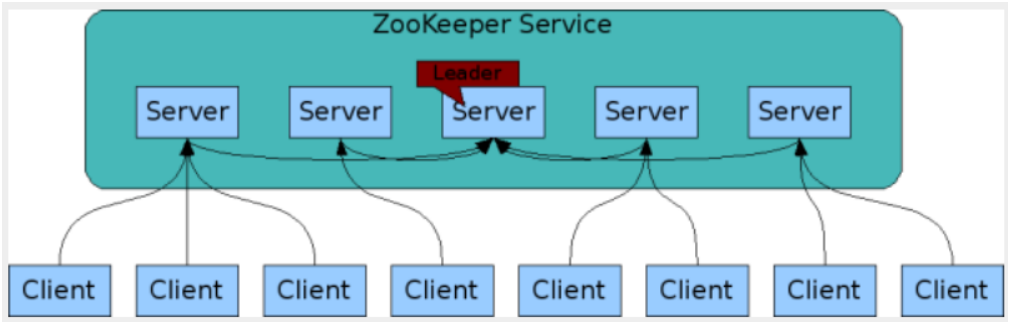


ภาพที่ 7.6 หลักการทำงานของสคูป (ที่มา : https://pocfarm.wordpress.com/2016/04/26/sqoop-sql-to-hadoop/)

ภาพที่ 7.6 สคูปเป็นตัวกลางในการนำข้อมูลจากระบบฐานข้อมูลเชิงสัมพันธ์มาแปลงเป็นแบบระบบแฟ้มข้อมูลแบบกระจายฮาดูป และเป็นตัวกลางระหว่างการนำข้อมูลออกจากระบบแฟ้มข้อมูลแบบกระจายฮาดูปให้เป็นระบบฐานข้อมูลเชิงสัมพันธ์

การนำสคูปไปประยุกต์ใช้งานในด้านการนำข้อมูลนำเข้าจากระบบเดิมที่เก็บข้อมูลเป็นแบบระบบฐานข้อมูล เชิงสัมพันธ์มาเป็นแบบระบบแฟ้มข้อมูลแบบกระจายฮาดูป เพื่อเพิ่มประสิทธิภาพในการประมวลผล หรือในการปรับเปลี่ยนระบบการทำงานแบบเดิมมาเป็นระบบการประมวลผลข้อมูลขนาดใหญ่ ตัวอย่างระบบงาน ได้แก่ ระบบงานธนาคารสาขาที่โอนข้อมูลจากสาขาที่เก็บข้อมูลแบบระบบฐานข้อมูลเชิงสัมพันธ์มาที่สำนักงานใหญ่เพื่อประมวลผลและวิเคราะห์ข้อมูลเชิงลึกบนข้อมูลขนาดใหญ่ ได้แก่การหาผลสรุปยอดรวมการทำธุรกรรมทางการเงินในแต่ละวันของทุกสาขา แยกตามจังหวัดในรูปของแผงควบคุม (Dashboard)

1.3 ซูคีปเปอร์ (Zookeeper) มีสัญลักษณ์ของซอฟต์แวร์เป็นรูปคน สื่อความหมายถึงผู้ดูแลสวนสัตว์หรือ ทำหน้าที่จัดการประสานการทำงานระหว่างซอฟต์แวร์ต่างๆ ภายในระบบนิเวศที่เกี่ยวข้องของฮาดูป โดยซอฟต์แวร์ ซูคีปเปอร์ทำหน้าที่จัดการเรื่องการบริการ การประสานงานเกี่ยวกับกระบวนการทำงานของแอปพลิเคชันแบบกระจายและให้บริการทำซ้ำข้อมูล (Data Replication) ไปยังเครื่องแม่ข่ายหรือเซิร์ฟเวอร์ (Server) และลูกข่ายหรือไคลเอนต์ (Client) ในระบบ เช่นในกรณีที่เครื่องใดเครื่องหนึ่งในระบบ เกิดความผิดพลาดจะมีการทำการติดตามและตรวจสอบ (Tracing and Monitoring) จากนั้นจะประสานงานเพื่อนำข้อมูลจากเครื่องอื่นในระบบเครือข่ายที่ได้ทำซ้ำข้อมูลมาใช้ทันที เพื่อให้ระบบการทำงานสามารถทำงานได้ตามปกติ ต่อเนื่องและไม่หยุดชะงัก ทำให้เกิดความน่าเชื่อถือว่า แอปพลิเคชันนั้นๆ จะสามารถทำงานได้อย่างถูกต้องถึงแม้เครื่องใดเครื่องหนึ่งในระบบเกิดปัญหาหรือขัดข้อง และมีความสามารถรองรับการทำงานพร้อมๆ กันในการประมวลผลข้อมูล ดังภาพที่ 7.7

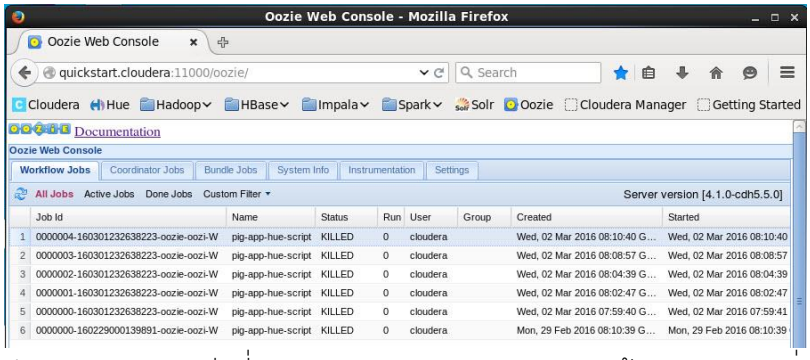


ภาพที่ 7.7 หลักการทำงานของซูคีปเปอร์ (ที่มา : https://cwiki.apache.org)

ภาพที่ 7.7 ซูคีปเปอร์จะดำเนินการกระจายข้อมูลโดยแบ่งการทำงานเป็น 2 ส่วน ได้แก่ Server และ Client โดยที่เซิร์ฟเวอร์ที่มีการกระจายข้อมูลจริงและมีการติดต่อระหว่างกันเพื่อให้ไคลเอนต์สามารถเชื่อมต่อกับเซิร์ฟเวอร์ ใดๆ ใน Cluster เพื่อให้ได้รับผลจากการทำงานที่เหมือนกันบนระบบการประมวลผลข้อมูลแบบกระจาย

ตัวอย่างลักษณะงานที่นำซูคีปเปอร์ไปประยุกต์ใช้ได้แก่องค์กรหรือบริษัท E-Commerce หรือ Mobile Commerce ที่มีการเก็บข้อมูลของสินค้าและบริการต่างๆ ไว้ที่เซิร์ฟเวอร์ขององค์กรหรือบนระบบ Cloud ที่ประกอบด้วยเซิร์ฟเวอร์หลายๆ ตัวเชื่อมโยงกัน โดยใช้ซูคีปเปอร์ ในการเชื่อมประสานงานเกี่ยวกับกระบวนการทำงานของแอปพลิเคชันแบบกระจายเพื่อประมวลผลข้อมูลใหญ่และให้บริการ Replicate ข้อมูลจากเซิร์ฟเวอร์อื่นๆ เพื่อป้องกันการผิดพลาด ในกรณีที่เซิร์ฟเวอร์ใดเซิร์ฟเวอร์หนึ่งเกิดปัญหาไม่สามารถทำงานได้ ก็สามารถเปลี่ยนเซิร์ฟเวอร์ทันทีและนำข้อมูลที่ถูกทำซ้ำบนเซิร์ฟเวอร์อื่นมาใช้งานแทนได้ เพื่อเป็นหลักประกันว่าแอปพลิเคชันนั้นสามารถทำงานได้อย่างถูกต้องต่อไป

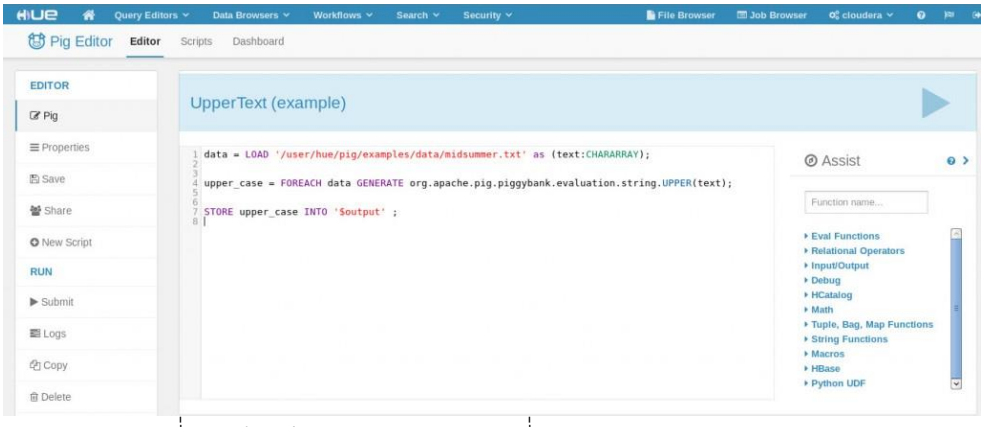
1.4 อูซี (Oozie) มีสัญลักษณ์ของซอฟต์แวร์เป็นตัวอักษรคำว่า oozie สีเหลือง โดยซอฟต์แวร์อูซีทำหน้าที่จัดการเรื่องเครื่องมือที่ใช้ทำ Workflow และสร้างเป็น Job Scheduler เพื่อจัดการงานของ Apache Hadoop ดังภาพที่ 7.8



ภาพที่ 7.8 ตัวอย่างการทำงานของอูซี (ที่มา : http://www.autosoft.in.th/ห้องทดลอง1/มาเริ่มต้นศึกษา-big-data-hadoop-กันกั/)

จากภาพที่ 7.8 ตัวอย่างการทำงานของอูซีในการแสดงสถานะการทำงานของแต่ละ Job ตามลำดับการทำงาน โดยอูซีถูกนำไปประยุกต์การทำงานเกี่ยวกับการจัดการ Job ของ JAVA MapReduce, Streaming MapReduce, Pig, Hive, Sqoop และ DistCp (distributed copy) เป็นเครื่องมือที่ใช้ในการสำเนา (Copy) ข้อมูลภายในหรือระหว่าง Cluster

1.5 พิก (Pig) มีสัญลักษณ์ของซอฟต์แวร์เป็นรูปหมู โดยซอฟต์แวร์พิกทำหน้าที่อำนวยความสะดวกในการเขียนคำสั่งแบบภาษาสคริปต์ (Script) ซึ่งมีตัวดำเนินการ (Operator) เข้ามาช่วยในการเขียนคำสั่งเช่น Join, Filter, Sort หรือ Group ที่ช่วยให้ประมวลผลข้อมูลโดยไม่ต้องเขียนโปรแกรมเพื่อดำเนินการแมปรีดิวซ์ด้วยภาษาจาวา (JAVA) ที่จะต้องเขียนคำสั่งหลายๆ บรรทัด เพื่อให้ทำงาน แต่การเขียนคำสั่งของพิกจะเขียนกระชับ ง่ายและโดยไม่ต้องเขียนหลายบรรทัด ซึ่งจะเรียกภาษาแบบนี้ว่าพิกลาติน (Pig Latin) ดังภาพที่ 7.9



ภาพที่ 7.9 ตัวอย่างการทำงานของพิก (ที่มา : https://pig.apache.org/)

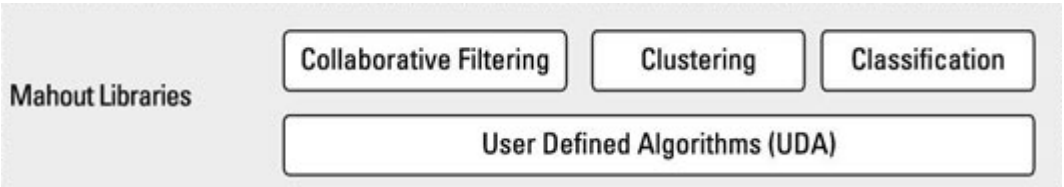
จากภาพที่ 7.9 คำสั่งการทำงานของพิกอาทิ คำสั่ง LOAD เป็นการนำข้อมูลมาเก็บไว้ เช่น

data = LOAD '/user/hue/pig/examples/data/midsummer.txt' as (text:CHARARRAY);

เป็นการนำข้อมูลจากที่เก็บข้อมูลของ HDFS ที่ตำแหน่ง /user/hue/pig/examples/data/ โดยนำข้อมูลจากไฟล์ชื่อ midsummer.txt ซึ่งเป็นการเก็บข้อมูลในแบบ Array ไปไว้ในตัวแปรที่ data

การนำพิกไปประยุกต์ใช้งานในด้านการทำอีทีแอล (Extract-Transform-Load : ETL) หมายถึงกระบวนการนำข้อมูลนำเข้าสู่ระบบคลังข้อมูล (Data Warehouse หรือ DW) โดยระบบที่ออกแบบเอาไว้จะนำข้อมูลนำเข้ามาจากหลายๆ ที่ เพื่อทำการตรวจสอบคุณภาพและปรับข้อมูลให้เป็นไปในรูปแบบเดียวกันเพื่อให้ข้อมูลจากหลายๆ แหล่ง สามารถใช้งานร่วมกันได้และท้ายที่สุดทำการส่งมอบ (Delivery) ข้อมูลเหล่านั้นในรูปแบบที่ง่ายต่อการใช้งาน เช่นกราฟ หรือรายงานสรุปผล เป็นต้น เพื่อใช้ในการตัดสินใจของผู้บริหารหรืออำนวยความสะดวกในการวิเคราะห์ข้อมูลขององค์กร ตัวอย่างระบบงานได้แก่ระบบงานธนาคาร บริษัทผู้ให้บริษัทเว็บโฮสติ้งหรือตลาดหลักทรัพย์

1.6 มาฮาวท์ (Mahout) มีสัญลักษณ์ของซอฟต์แวร์เป็นรูปควาญช้างนั่งอยู่บนหลังช้างฮาดูป สื่อความหมายถึงการควบคุมการประมวลผลข้อมูลขนาดใหญ่บนฮาดูป โดยซอฟต์แวร์มาฮาวท์ทำหน้าที่จัดการเรื่องการวิเคราะห์ข้อมูลเชิงทำนายผล เช่นงานทางด้านวิทยาศาสตร์ข้อมูล (Data Science) ที่นำข้อมูลใหญ่มาทำการวิเคราะห์หรืองานวิจัยต่างๆ ที่เกี่ยวกับข้อมูลขนาดใหญ่บนฮาดูป ซึ่งอาปาเชมาฮาวท์จะมีหลายๆ อัลกอริทึมที่รองรับการทำงานได้แก่ ระบบแนะนำข้อมูลแบบการกรองแบบร่วมมือกัน การจำแนกข้อมูล การจัดกลุ่มข้อมูล และ การสร้างแบบจำลองหัวข้อ ดังภาพที่ 7.10



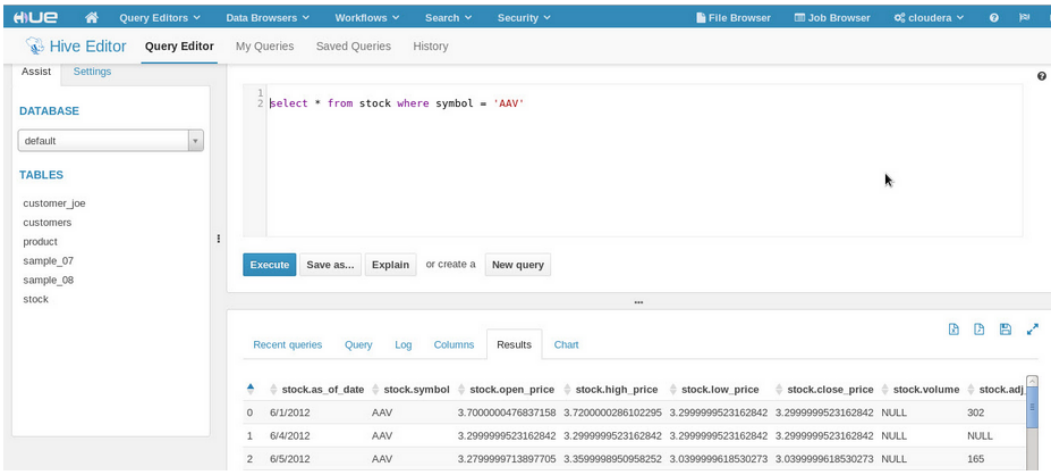
ภาพที่ 7.10 อัลกอริทึมที่รองรับการทำงานของมาฮาวท์ (ที่มา : http://www.dummies.com/programming/big-data/hadoop/machine-learning-with-mahout-in-hadoop/)

จากภาพที่ 7.10 อัลกอริทึมที่รองรับการทำงานของมาฮาวท์ได้แก่ การกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) การจัดกลุ่มข้อมูล (Clustering) การจำแนกข้อมูล (Classification) และการกำหนดหรือสร้างอัลกอริทึมขึ้นมาใช้เอง (User Defined Algorithms : UDA)

การนำมาฮาวท์ไปประยุกต์ใช้งานในด้านการทำเหมืองข้อมูล การให้คำแนะนำ เช่นระบบคำแนะนำ (Recommendation System) การซื้อสินค้าแบบพาณิชย์อิเล็กทรอนิกส์ (E-Commerce) เช่น eBay.com หรือ Amazon.com ซึ่งมีการทำเทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) มาประยุกต์การทำงานบนข้อมูลขนาดใหญ่ จำนวนมาก ตัวอย่างระบบงานได้แก่ ระบบบัตรเครดิต ระบบการคัดลอกวรรณกรรม ระบบคัดกรองอิเมล์ (Email) หรือระบบการพยากรณ์อากาศ

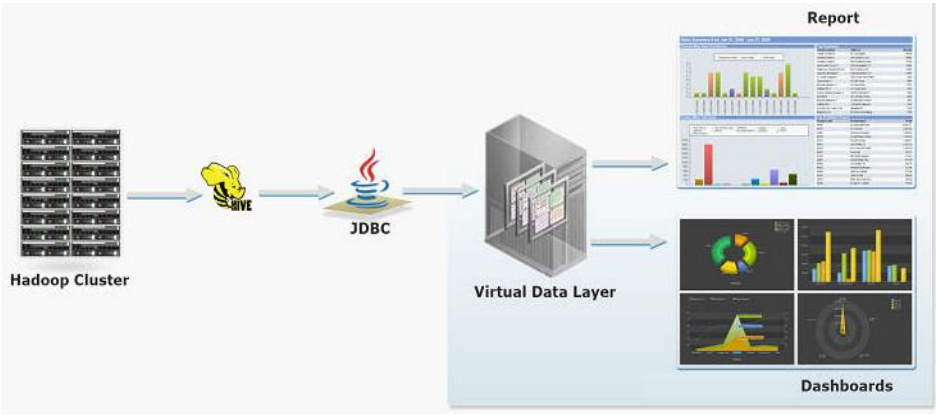
1.7 อาร์ (R Connectors) โดยซอฟต์แวร์อาร์ทำหน้าที่จัดการเรื่องการคำนวณทางสถิติและแสดงกราฟฟิกโดยเฉพาะบนฮาดูป

1.8 ไฮฟ (HIVE) มีสัญลักษณ์ของซอฟต์แวร์เป็นรูปหัวช้างสีเหลืองอยู่ในรังผึ้ง สื่อความหมายถึงคลังข้อมูล โดยซอฟต์แวร์ไฮฟทำหน้าที่จัดการเรื่องคลังข้อมูล (Data Warehouse) บนข้อมูลขนาดใหญ่ เพื่ออำนวยความสะดวกในการสืบค้นข้อมูล (Query) และการจัดการชุดข้อมูลขนาดใหญ่ โดยไฮฟมีรูปแบบภาษาการสืบค้นข้อมูลเหมือนภาษาเอสคิวแอล (SQL) จึงเรียกว่าไฮฟคิวแอล (HiveQL) ดังภาพที่ 7.11



ภาพที่ 7.11 ตัวอย่างการทำงานของไฮฟ

จากภาพที่ 7.11 แสดงหน้าจอการเขียนคำสั่งด้วยภาษา HiveQL คือคำสั่ง “Select * from stock where symbol = ‘AAV’ เพื่อทำการแสดงข้อมูลทั้งหมด จากตาราง stock ที่มีเงื่อนไขว่า symbol มีค่าเท่ากับข้อความว่า ‘AAV’ ซึ่งคำสั่งนี้จะเหมือนกับภาษา SQL โดยไม่ได้ดึงข้อมูลจากระบบฐานข้อมูล แต่ดึงข้อมูลจากข้อมูลที่มีการจัดเก็บเป็นแบบระบบแฟ้มข้อมูลแบบกระจายฮาดูป หรือ HDFS โดยที่ไม่ต้องเขียนคำสั่งการทำงานแบบ MapReduce บนHDFS เนื่องจาก Hive จะทำการแปลง HiveQL เป็น MapReduce แล้วทำการประมวลผลคำสั่งเป็นแบบ Batch เพื่อให้สามารถเข้าถึงข้อมูลโดยผ่านทาง ODBC/JDBC ดังภาพที่ 7.12

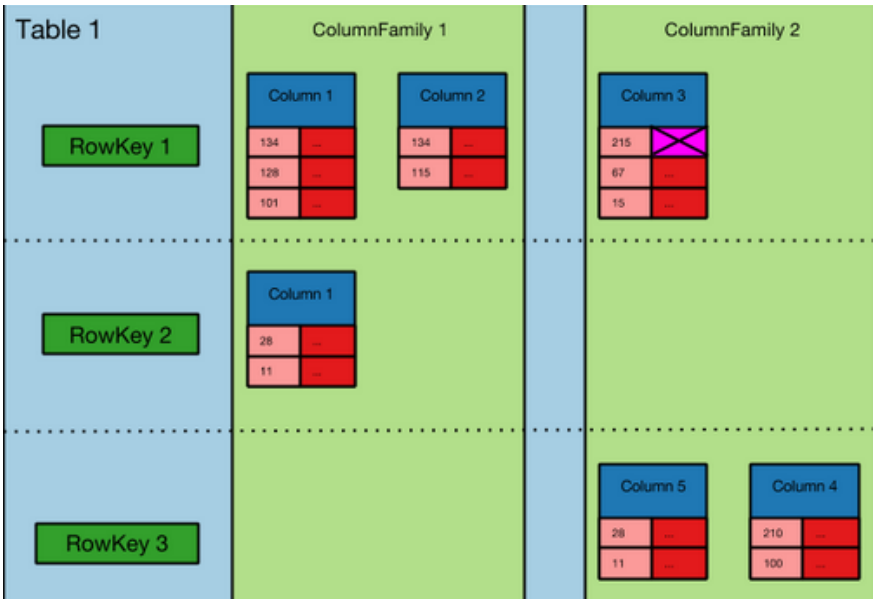


ภาพที่ 7.12 หลักการทำงานของไฮฟ (ที่มา : https://hive.apache.org/)

จากภาพที่ 7.12 เป็นการนำข้อมูลจาก Hadoop Cluster ที่มีการจัดการข้อมูลเป็นแบบระบบแฟ้มข้อมูลแบบกระจาย ฮาดูป เพื่อกระจายข้อมูลไปประมวลผลใน Node ต่างๆ ของ cluster จากนั้น Hive จะทำการแปลง HiveQL และทำการเชื่อมโยงข้อมูลที่อยู่ใน Hadoop Cluster ด้วยเจดีบีซี (JDBC ย่อมาจาก Java Database Connectivity คือ API (Application Programming Interface) หรือ Library ในจาวาที่ใช้สำหรับติดต่อกับฐานข้อมูลที่เป็นแบบ Relational อย่างเช่น MS SQL, Oracle, MySQL, DB2 และ Informix เป็นต้น โดย JDBC จะช่วยทำให้สามารถเพิ่ม แก้ไข ลบ หรือเรียกดูข้อมูลที่เก็บไว้ในฐานข้อมูลจากโปรแกรมจาวาที่เขียนขึ้นได้หรืออาจเรียกว่าเป็นตัวเชื่อมต่อระหว่างโปรแกรมกับฐานข้อมูลของจาวา) และแสดงผลข้อมูลออกมาเป็นรายงาน (Report) หรือแผงควบคุม (Dashboard) ที่สวยงามและเข้าใจได้ง่าย

การนำไฮฟไปประยุกต์ใช้งานในด้านการทำเหมืองข้อมูลหรือธุรกิจอัจฉริยะ (Business Intelligence: BI) คือการรวบรวมข้อมูล การจัดเก็บข้อมูล การวิเคราะห์และการเข้าถึงข้อมูล รวมถึงการเรียกดูข้อมูลในหลากหลายมุมมอง (Multidimensional Model) ของแต่ละหน่วยงาน ซึ่งช่วยให้ผู้ใช้งานในองค์กรทำการตัดสินใจทางธุรกิจที่ดียิ่งขึ้นบนข้อมูลขนาดใหญ่จำนวนมาก ตัวอย่างลักษณะงานที่นำไฮฟไปประยุกต์ใช้ได้แก่การพัฒนาและวางระบบงานสารสนเทศทางด้านธุรกิจเช่าซื้อรถยนต์-เงินฝาก ให้กับธนาคาร ในการจัดการข้อมูลเชิงลึก ครอบคลุมการรายงานผล ในลักษณะเวลาจริง (Real time) หรือแสดงผลแบบทันทีทันใด และช่วยให้ธนาคารสามารถพัฒนาสินค้าและบริการได้ตรงความต้องการของลูกค้ามากยิ่งขึ้น หรือในการประมวลผลในเชิงวิเคราะห์แบบออนไลน์ (OLAP) ในการนำเสนอข้อมูลในรูปแบบของลูกบาศก์ (CUBE) ในอุตสาหกรรมอาหารสำเร็จรูป ที่ช่วยให้ผู้บริหารได้รับข้อมูลรายงานที่รวดเร็วและตอบสนองความต้องการของผู้ใช้งาน

1.9 เอชเบส (HBase)) มีสัญลักษณ์ของซอฟต์แวร์เป็นตัวอักษรคำว่า HBASE สีแดง โดยซอฟต์แวร์เอชเบสทำหน้าที่จัดการเรื่องการอ่านและเขียนข้อมูลแบบเวลาจริง (Realtime) โดยจะนำข้อมูลมาเก็บในรูปแบบตารางใหญ่ (Big Table) คือการเก็บข้อมูลที่ไม่จำกัดจำนวนแถว (Rows) หรือคอลัมน์ (Columns) ในตาราง ซึ่งอาปาเชเอชเบสจะเป็นเสมือนการทำให้ฮาดูปเป็น NoSQL Database แบบสมบรูณ์ ดังภาพที่ 7.13

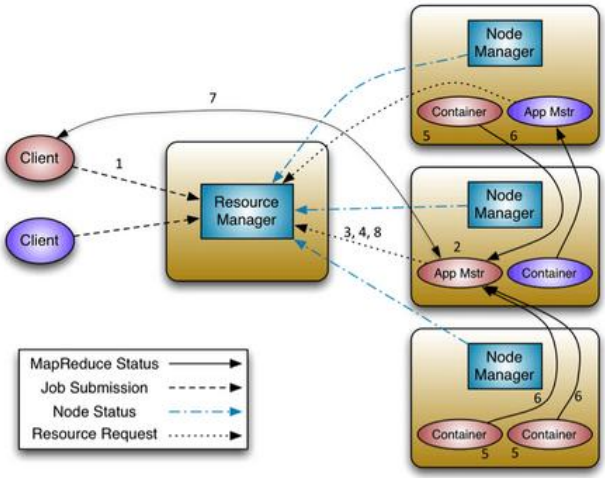


ภาพที่ 7.13 หลักการทำงานของเอชเบส (ที่มา :https://hbase.apache.org/)

จากภาพที่ 7.13 จะนำข้อมูลในรูปแบบตารางหรือ Table 1 มากำหนด RowKey ตามลำดับเช่น 1,2,..., N และกำหนด ColumnFamily เป็นตารางของ Column ตามลำดับ การดำเนินการแบบนี้เรียกว่า Column-Oriented ใน NOSQL (ซึ่งรายละเอียดได้กล่าวไว้แล้วในหน่วยที่ 4)

การนำเอชเบสไปประยุกต์ใช้งานในด้านการประมวลผลแบบเวลาจริง (Realtime) หรือทันทีทันใดเช่น การประมวลผลข้อมูลของบริษัท ที่รองรับการสอบถามจากลูกค้ามากกว่า 30,000 ครั้งต่อวินาที ที่ต้องการคำตอบแบบทันทีทันใด ณ เวลานี้ หรือข้อมูลการวิเคราะห์ผลตลาดหลักทรัพย์แบบเวลาจริง ตัวอย่างระบบงานได้แก่ ระบบบริษัทสายการบิน ระบบโรงพยาบาล ระบบตลาดหลักทรัพย์ ระบบโลจิสติกส์ หรือระบบประมวลผลภาพถ่ายดาวเทียม

1.10 ยาน (YARN MapReduce v2) มาจากคำว่า YARN (Yet Another Resource Negotiator) หรือ Hadoop MapReduce รุ่นที่ 1 ดังภาพที่ 7.14



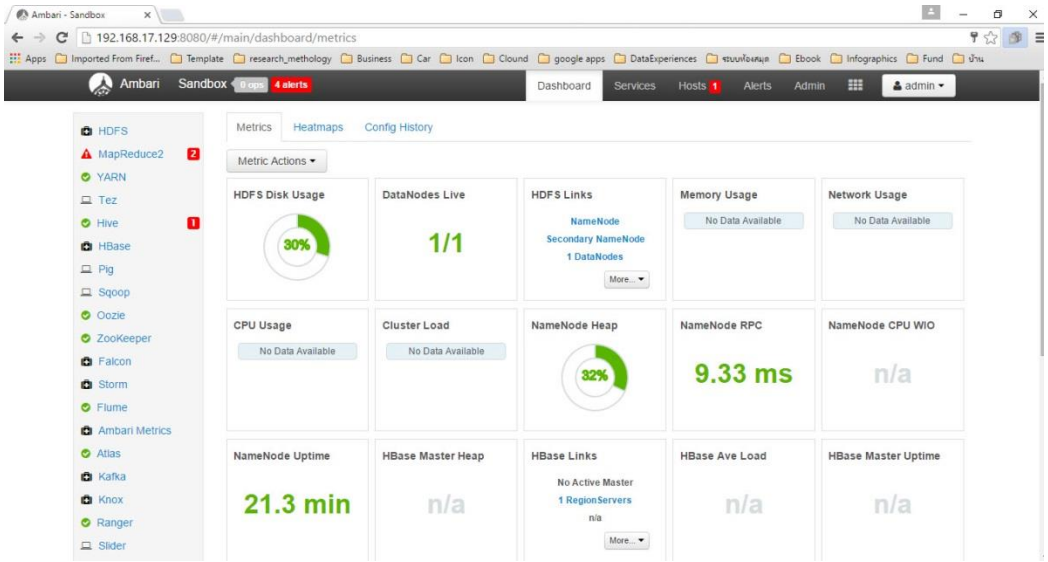
ภาพที่ 7.14 หลักการทำงานของยาน (ที่มา : https://hortonworks.com/blog/apache-hadoop-yarn-concepts-and-applications/)

จากภาพที่ 7.14 YARN จะประกอบไปด้วยส่วนประกอบหลัก 2 ส่วนคือ Resource manager และ Node manager โดยที่ Resource manager จะทำหน้าที่ค่อยดูแล Resource ของทั้ง Cluster ซึ่งจะมีแค่ 1 ตัวต่อ Cluster เท่านั้น ส่วน Node manager จะทำหน้าที่ดูแลจัดการและ Monitor resource ในแต่ละ Node หรือ Contrainer รวมถึง Launch โปรแกรมที่จะมาใช้ Resource ใน Node นั้นๆ และติดต่อไปยัง Resource manager ที่ดูแลระดับ Cluster โดย Node manager นี้จะถูกติดตั้งตามจำนวน Node ใน Cluster

สำหรับการจัดการแต่ละแอปพลิเคชันนั้น ส่วน Client ของ User ต้องติดต่อกับทาง Resource manager เพื่อให้ทาง Resource manager จัดการหา Node manager ที่มีความที่จะรับ Application master ไป Launch ที่ node ได้ โดย Application master อาจจะเป็น Application ที่ Run จบแล้วส่ง Result ของการ Process กลับไปยัง Client ได้เลยหรืออาจจะเป็น Application ที่ต้องการการ Process ที่ซับซ้อนและต้องการ Resource จาก Resource manager เพิ่มระหว่างการ Process ได้อีกด้วย

1.11 แฟ้มข้อมูลแบบกระจายฮาดูป (Hadoop Distributed File System หรือ เอชดีเอฟเอส HDFS) ที่ทำหน้าที่เป็นส่วนเก็บข้อมูลซึ่งจะเก็บข้อมูลขนาดใหญ่ที่จะแบ่งเป็นไฟล์ย่อยขนาดใหญ่เก็บลงใน Data Node จำนวนมาก โดยจะมี Master Node ที่ทำหน้าที่ระบุตำแหน่งของข้อมูลที่เก็บใน Data node (ซึ่งรายละเอียดได้กล่าวแล้วในหน่วยที่ 5)

1.12 แอมบารี (Ambari) โดยซอฟต์แวร์แอมบารีทำหน้าที่จัดการเรื่องหน้าจอ Web UI ของค่าย ฮอนทรอนเวิร์ก (Hortonworks) ทำให้สามารถจัดการ Hadoop ผ่านเว็บเบราว์เซอร์ทำได้ง่ายและสะดวก ดังภาพที่ 7.15



ภาพที่ 7.15 ตัวอย่างการทำงานของแอมบารี (ที่มา : https://blog.codecentric.de/en/2014/04/hadoop-cluster-automation/)

จากภาพที่ 7.15 ตัวอย่างการทำงานของแอมบารีผ่านเว็บเบราว์เซอร์ โดยแสดงรายละเอียด Disk usage สถานะการทำงานของ DataNodes Live จำนวน HDFS Links การใช้ Memory หรือ Network usage เพื่อทำให้การจัดการ Hadoop เป็นเรื่องที่ง่ายและสะดวกมากยิ่งขึ้น

ประมวลความรู้ของหลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร) (ข้อบังคับ)

- การนำความรู้และทักษะที่ได้เรียนรู้จากเนื้อหาแต่ละวิชาในหลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร) ไปประยุกต์ใช้ ตัวอย่างประกอบ และการเขียนตอบในเชิงบูรณาการ

แนวทางเตรียมสอบ 09600 ประมวลความรู้ วท.ม.ไอซีที ภาค 2/2563

หลักสูตรปรับปรุง พ.ศ. 2554 (นักศึกษารหัส 56-59)

ภาคทฤษฎี: (3 ข้อ: ทำทุกข้อ)

99701 ธุรกิจอิเล็กทรอนิกส์และการประยุกต์

ทำการศึกษาเกี่ยวกับ Digital Transformation สำหรับการพัฒนาและปรับปรุงธุรกิจ เพื่อให้สอดรับกับการเปลี่ยนแปลงของยุคดิจิตัล 4.0 รวมถึงเทคโนโลยีต่าง ๆ ที่เกี่ยวข้อง

99702 การพัฒนาระบบสารสนเทศ การบริหารโครงการและการประยุกต์

- ให้ศึกษาเนื้อหาชุดวิชาและกิจกรรมอีเลินนิ่ง เรื่อง การเขียน use case diagram และ activity diagram
- ให้ทบทวนและฝึกเขียน Use Case Diagram และ Activity Diagram การสร้างแผนภาพคลาส (Class Diagram) การแสดง classes, attributes, operations, associations /aggregations และ generalizations จากกรณีศึกษาที่กำหนดให้ โดยใช้กิจกรรมอีเลินนิงเรื่อง ระบบ Banking application เป็นแนวทางในการศึกษา

99703 การจัดการเทคโนโลยีสารสนเทศและการสื่อสารเชิงกลยุทธ์

- สไลด์ประกอบการบรรยายของอ.ดร.เฉลิมศักดิ์ เลิศวงศ์เสถียร เรื่อง “Information Technology Strategic Planning”
- หน่วยที่ 13 การวางแผนกลยุทธ์เทคโนโลยีสารสนเทศ เรื่องที่ 13.2.1 การวิเคราะห์สภาพแวดล้อม และตารางที่ 13.2 การวิเคราะห์แรงแข่งขันทั้ง 5

ภาคประยุกต์: (3 ข้อ: บังคับทำข้อสอบ 1) 99705 2) ข้อสอบบูรณการ และ 3) เลือกทำ 99704 หรือ 99707)

99705 ความมั่นคงด้านเทคโนโลยีสารสนเทศและการสื่อสาร (ข้อบังคับ)

- ศึกษาประเภทและรูปแบบการโจมตีระบบเครือข่ายคอมพิวเตอร์ลักษณะต่างๆ ได้ พร้อมยกตัวอย่างและเสนอแนะแนวทางการป้องกัน

<u>เลือกทำ</u>

99704 คลังข้อมูล เหมืองข้อมูล และธุรกิจอัจฉริยะ.

- การประยุกต์คลังข้อมูล เหมืองข้อมูล และธุรกิจอัจฉริยะ ในอุตสาหกรรม/งานต่างๆ และลักษณะของโจทย์ปัญหา ที่เหมาะสมหรือควรใช้ คลังข้อมูล เหมืองข้อมูล และธุรกิจอัจฉริยะ ในการแก้ปัญหา
- ขั้นตอนการทำเหมืองข้อมูล และอัลกอริธึมในการทำเหมืองข้อมูล ที่เหมาะสมในการแก้ปัญหา
- แบบจำลองเชิงแนวคิด (conceptual model/data mart model/ multi-dimensional model) และรายงานโอแลป (OLAP) ของระบบธุรกิจอัจฉริยะ

99707 ระบบสารสนเทศภูมิศาสตร์และการประยุกต์

ภูมิสารสนเทศเป็นศาสตร์ที่ครอบคลุมถึงระบบสารสนเทศภูมิศาสตร์ ภูมิสารสนเทศ จึงเป็นวิทยาศาสตร์และเทคโนโลยีที่เกี่ยวข้องกับการได้มา (Capture) การบูรณาการ (Integrating) การวิเคราะห์ (Analyzing) การจัดการ (Managing) และ การตีความ (Depicting) ข้อมูลข่าวสารเชิงพื้นที่ อันประกอบไปด้วยข้อมูล 3 ด้าน คือ 1) ทำเลที่ตั้ง (Location) ที่สามารถบอกเป็นค่าพิกัดที่แน่นอนได้ 2) สภาพแวดล้อมทางธรรมชาติเป็นข้อมูลที่แสดงถึงสิ่งแวดล้อมที่เกิดขึ้นเองตามธรรมชาติและ 3) สภาพแวดล้อมทางวัฒนธรรม เป็นข้อมูลที่แสดงถึงสิ่งแวดล้อมที่มนุษย์สร้างขึ้น

ประเภทข้อมูลในระบบสารสนเทศภูมิศาสตร์ประกอบด้วยข้อมูล 2 ประเภท คือ 1)ข้อมูลเชิงพื้นที่ (Spatial data)เป็นข้อมูลที่เกี่ยวข้องกับตำแหน่งที่ตั้งของข้อมูลต่าง ๆ บนพื้นโลก สามารถอ้างอิงกับตำแหน่งทางภูมิศาสตร์แสดงสัญลักษณ์ได้ 3 รูปแบบ คือ จุด เส้น และพื้นที่ 2) ข้อมูลเชิงคุณลักษณะ (Non-ซึ่งเป็นลักษณะข้อมูลเชิงบรรยาย โดยจะอธิบายถึงคุณลักษณะต่าง ๆ ในพื้นที่นั้น ให้ศึกษารายละเอียดเพิ่มเติม

ความแตกต่างระหว่างข้อมูลของระบบสารสนเทศทั่วไปกับข้อมูลของระบบสารสนเทศภูมิศาสตร์ รวมถึงการใช้ประโยชน์เป็นสิ่งที่นักศึกษาต้องไปศึกษารายละเอียดเพิ่มเติม

ฐานข้อมูลระบบสารสนเทศภูมิศาสตร์มีลักษณะการเชื่อมโยงระหว่างฐานข้อมูลเชิงบรรยายกับข้อมูลเชิงเส้นหรือข้อมูลเชิงภาพ กระบวนการในการวิเคราะห์ข้อมูลเชิงพื้นที่ มีกระบวนการหลัก คือ การกำหนดจุดประสงค์ การวางแผนวิเคราะห์รวบรวมข้อมูล ทำการวิเคราะห์ และสรุปผล

กระบวนการในการวิเคราะห์ ประกอบด้วย **1.กำหนดจุดประสงค์** หรือระบุปัญหาที่ต้องการหาคำตอบให้ชัดเจนจัดลำดับความสำคัญของจุดมุ่งหมายให้ชัดเจน (set priority) ว่าอะไรคือจุดประสงค์หลัอะไรคือจุดประสงค์รอง สิ่งใดมีความสำคัญหรือสิ่งใดที่เป็นเพียงผลพลอยได้จากการวิเคราะห์ **2.วางแผนการวิเคราะห์และรวบรวมข้อมูล** เป็นการศึกษาและรวบรวมปัจจัยในการวิเคราะห์รวมไปถึงสถิติลักษณะ คุณภาพและข้อจำกัดของข้อมูลที่มี การจัดกลุ่มข้อมูล เลือก modelที่เหมาะสมในการวิเคราะห์ประมาณงบประมาณที่จะใช้ ระบุความต้องการ ความเสี่ยง หรือรายละเอียดของข้อมูลที่จะช่วยในการตัดสินใจ และระยะเวลาการวิเคราะห์ ทรัพยากรที่ต้องใช้ ปัจจัยเสี่ยงและความเป็นไปได้ มาตราฐานแหล่งที่มาของข้อมูล การสุ่มตัวอย่าง ข้อจำกัดและการทบทวนปัญหาที่ต้องการคำตอบในขั้นต้น **3. ทำการวิเคราะห์** ในการวิเคราะห์ข้อมูลนั้นอาจมีความจำเป็นต้องแบ่งการวิเคราะห์ออกเป็นประเด็นย่อยๆ ตามจุดประสงค์ ขั้นตอน หรือmodel ที่ได้วางไว้ในตอนแรกเพื่อให้ง่ายในการวิเคราะห์ และเพื่อให้การวิเคราะห์นั้นมีความถูกต้องและชัดเจนมากที่สุด ผู้วิเคราะห์สามารถเลือกใช้โปรแกรมประยุกต์แบบสำเร็จรูปที่มี module หรือ function พร้อมใช้เพื่อความสะดวก หรือโปรแกรมแบบที่สามารถเขียน module หรือ function เพิ่มเติมเอง ได้ตามความต้องการ ไม่ว่าจะเป็น Arc GIS, Quantum GIS, GRASS GIS หรือโปรแกรม R4 **สรุปผล** กระบวนการสุดท้ายของกระบวนการวิเคราะห์ข้อมูลก็คือการสรุปผล ไม่ว่าจุดประสงค์หรือสมมุติฐานที่ที่ตั้งในเบื้องต้นนั้นจะได้รับคำตอบหรือไม่ อย่างไร เป็นไปตามความคาดหมายในเบื้องต้นหรือไม่ มีข้อดี ข้อด้อย ข้อจำกัดหรือขอบเขต คุณภาพ ความถูกต้อง และการประเมินขั้นตอนในการวิเคราะห์อย่างไร ผลลัพธ์ที่ได้ ยอมรับได้หรือไม่ คำแนะนำต่อไปเป็นอย่างไร ต้องทำการวิเคราะห์ต่อหรือไม่ **ให้นักศึกษาศึกษารายละเอียดเพิ่มเติมในฟังก์ชั่นของการวิเคราะห์ข้อมูล**

ประมวลความรู้ของหลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร) (ข้อบังคับ)

- การนำความรู้และทักษะที่ได้เรียนรู้จากเนื้อหาแต่ละวิชาในหลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร) ไปประยุกต์ใช้ ตัวอย่างประกอบ และการเขียนตอบในเชิงบูรณาการ