RMUTP

หลักสูตรการวิเคราะห์ข้อมูลขนาดใหญ่

Big Data Analytics

สถาบันสหวิทยาการ
ดิจิทัลและหุ่นยนต์
DIRI Digital Interdisciplinary and Robotics Institute

diri.rmutp diri.rmutp diri@rmutp diri_rmutp @diri.rmutp
http://diri.rmtup.ac.th
diri@rmutp.ac.th

สารบัญ

สารบัญรูป

สารบัญตาราง

Section 1: The basics of working with big data

1.1 วัตถุประสงค์การเรียนรู้

1) ทราบที่มาของข้อมูลขนาดใหญ่
2) ทราบความหมายของข้อมูลขนาดใหญ่
3) เข้าใจความท้าท้ายในโลกของข้อมูลขนาดใหญ่

1.2 ที่มาของข้อมูลขนาดใหญ่

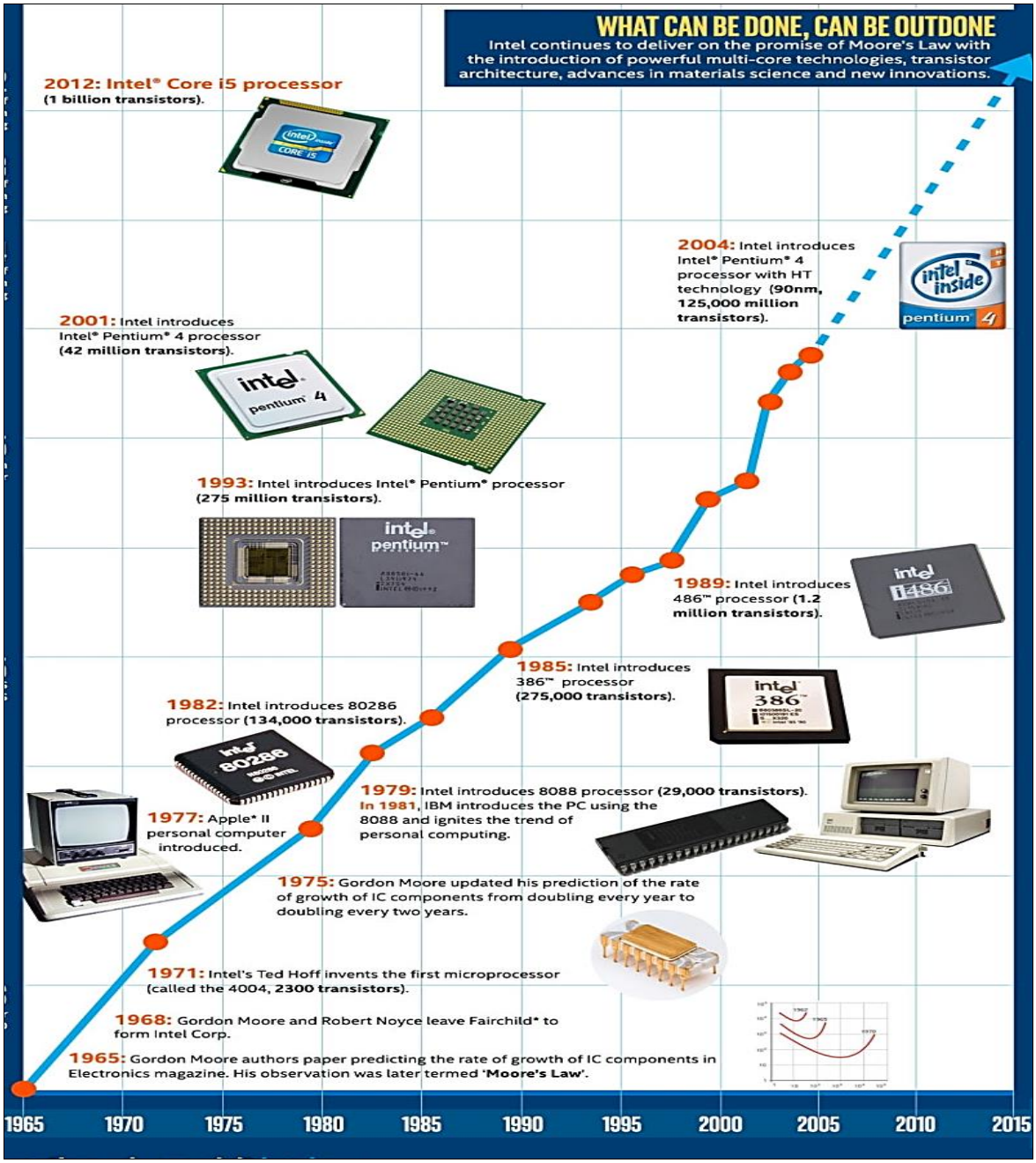
เทคโนโลยีข้อมูลขนาดใหญ่ (Big Data Technology) นั้นมีมานานแล้ว แต่อย่างไรก็ตาม เนื่องจากในยุคปัจจุบันนี้ เป็นยุคแห่งข้อมูลเชิงดิจิทัล ซึ่งทำให้ข้อมูลขนาดใหญ่นั้นมีความสำคัญกับการนำไปใช้ประโยชน์ในหลายๆ องค์กร และสามารถประยุกต์ใช้ได้ในหลายๆ ด้าน

กอร์ดอน มัวร์ (Gordon E. Moore) เป็นผู้ร่วมก่อตั้งบริษัทอินเทล ได้ใช้หลักการสังเกตตั้งกฎของมัวร์ (Moore's law) ขึ้น ซึ่งเขาบันทึกไว้ว่า ปริมาณของทรานซิสเตอร์บนวงจรรวม จะเพิ่มเป็นเท่าตัวทุก 18เดือน ซึ่งก็เป็นจริงตามนั้น หลังจากที่มีการค้นพบวงจรรวม (ไอซี) เพียงแค่ 4 ปี ดังรูปที่ 1.1

ความสามารถของอุปกรณ์อิเล็กทรอนิกส์มากมาย เป็นไปตามกฎของมัวร์ (Moore's law) อย่างเห็นได้ชัด เช่น

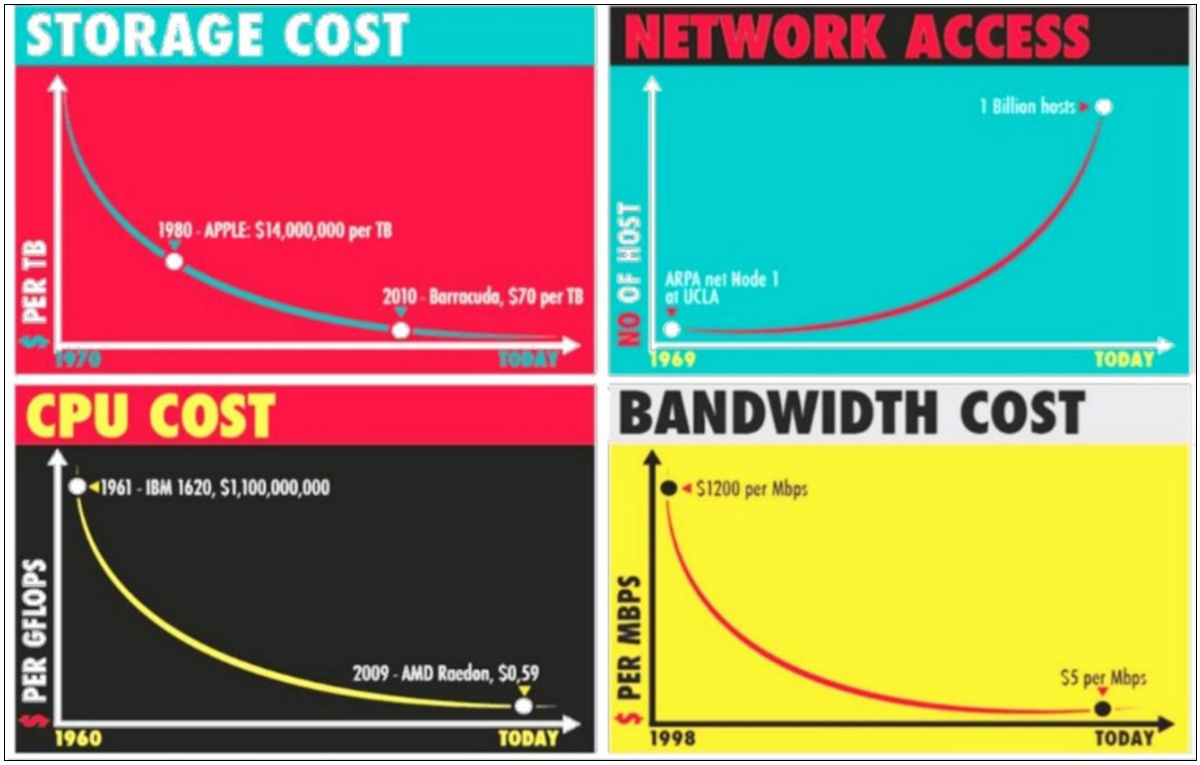
- ความเร็ว Computer Processor
- แบนด์วิชท์การสื่อสารและโทรคมนาคม
- หน่วยความจำของคอมพิวเตอร์
- ความจุฮาร์ดดิสก์

จากรูปที่ 1.2 ทำไมถึงเกิดคำว่าข้อมูลขนาดใหญ่ (Cearley, Burke, & Walker, 2016) แบนด์วิธการสื่อสารและโทรคมนาคมที่สูงขึ้นในราคาที่ถูกลงทำให้ประชาชนบุคคลทั่วไปเข้าถึง Internet มากขึ้นและทำให้เกิดข้อมูลมากมายในเครือข่าย Internet ราคา storage ที่ถูกลงมากทำให้สามารถเก็บข้อมูลได้ใน volume ที่มากขึ้น และที่สำคัญความสามารถของ CPU ที่สูงขึ้นในราคาที่ถูกลงทำให้ข้อมูลขนาดใหญ่ดังกล่าวสามารถประมวลผลได้ในเวลาที่สั้นลง ทั้งสามประเด็นคือเหตุผลของการเกิดคำว่า "ข้อมูลขนาดใหญ่" ในปัจจุบัน



รูปที่ 1.1 กฎของมัวร์

(Braun, 2015)



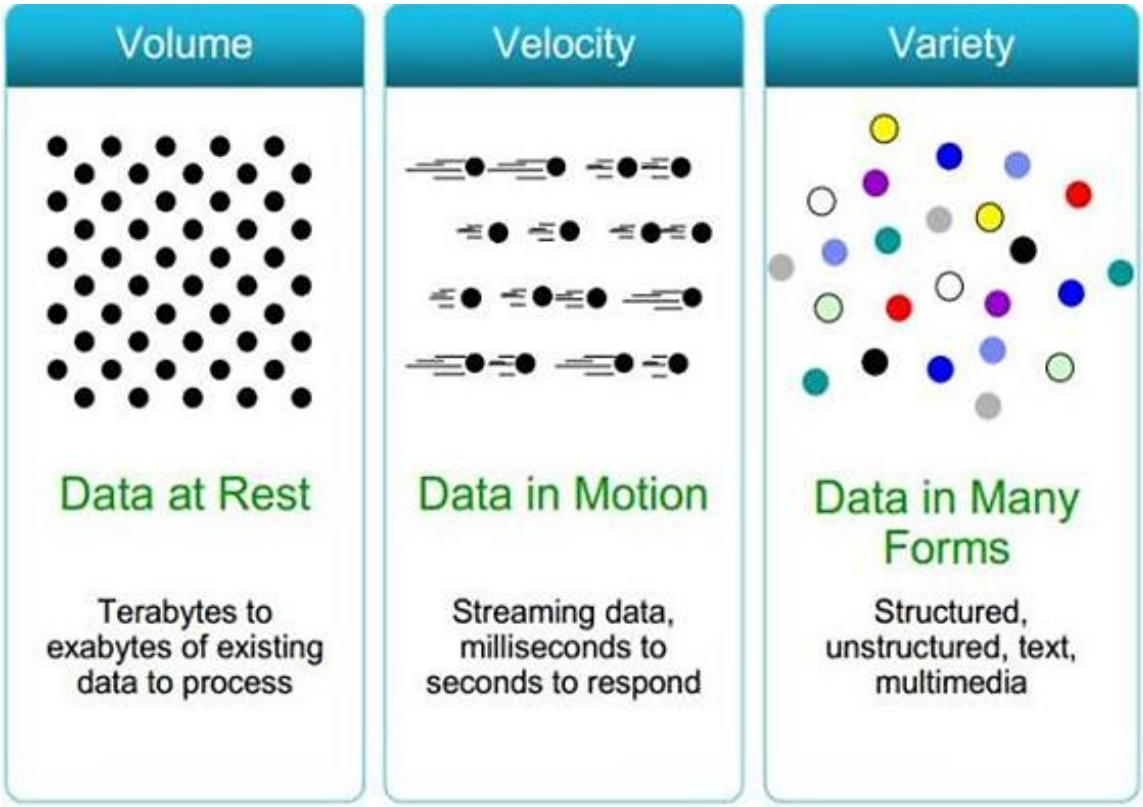
รูปที่ 1.2 ทำไมถึงเกิดคำว่าข้อมูลขนาดใหญ่

(Braun, 2015)

1.3 ข้อมูลขนาดใหญ่คืออะไร

คำว่า “ขนาดใหญ่” สำหรับแต่ละคนคงไม่เท่ากัน ลักษณะข้อมูลขนาดใหญ่มี 3V ดังนี้ Volume, Velocity และ Variety (Data Science Central, 2018) แสดงได้ดังรูปที่ 1.3

- Volume คือ ข้อมูลมีจำนวนมากเกินกว่าระบบฐานข้อมูลแบบเดิม ๆ จะสามารถที่จะจัดเก็บและจัดการได้
- Velocity คือ ข้อมูลจะมีการเกิดและเปลี่ยนแปลงตลอดเวลา เช่น ข้อมูลจาก Social Media ข้อมูลการซื้อขาย ข้อมูล Transaction การเงินหรือการใช้โทรศัพท์ และข้อมูลจาก Sensor ซึ่งเกิดขึ้นทุกวินาที
- Variety คือ ข้อมูลจะมีรูปแบบที่หลากหลายไม่ใช้ข้อมูลในรูปแบบตาราง Structure เพียงอย่างเดียวแต่รวมถึง Unstructured และ Semi structure เช่น Text, Image, VDO, XML และ JSON



รูปที่ 1.3 ลักษณะของข้อมูลขนาดใหญ่

(Braun, 2015)

1.4 ความท้าท้ายในโลกของข้อมูลใหญ่

เมื่อข้อมูลขนาดใหญ่เข้ามาเรื่อย ความสำคัญของวิทยาการข้อมูล (Data Sciences) ก็เพิ่มขึ้นอย่างรวดเร็วในการจัดการกับข้อมูลขนาดใหญ่ ข้อมูลขนาดใหญ่และสมัยใหม่นั้นไม่สามารถอาศัยฐานข้อมูลสัมพันธ์ (Relational Database) แบบสมัยก่อนได้อีกแล้ว เนื่องจากข้อมูลไม่ได้มีโครงสร้างชัดเจน (Unstructured Data) มีการเปลี่ยนแปลงอย่างรวดเร็ว ข้อมูลมีจำนวนมาก ไม่ได้มาจากการสุ่มตัวอย่างจากการสำรวจเหมือนในอดีต แต่ก็มีปัญหาไม่ยิ่งหย่อนไปกว่ากัน เนื่องจากมีปัญหาคุณภาพข้อมูล ข้อมูลทับซ้อนไม่ตรงกันหรือไม่สอดคล้องกัน มีข้อมูลเยอะแต่ก็มีข้อมูลสูญหาย (Missing data) มากมาย หน้าที่ของนักวิทยาศาสตร์ข้อมูล (Data Scientist) คือการจัดการรวบรวมข้อมูล วิเคราะห์ข้อมูล สรุปผล นำเสนอแนะ และนำสารสนเทศที่ได้ไปใช้งานให้เกิดประโยชน์สูงสุดแก่องค์กร

บทบาทหน้าที่นั้นนำมาซึ่งความท้าทายมากมายมาให้นักวิทยาศาสตร์ข้อมูล ต้องเรียนรู้ ปรับตัว และแก้ปัญหา

1. ข้อมูลมีขนาดใหญ่และไหลเข้ามารวดเร็วมาก จนต้องหา algorithm หรือขั้นตอนวิธีในการวิเคราะห์ให้เร็วขึ้น มีการแยกกันคำนวณ (Distributed computing)

2. ข้อมูลขนาดใหญ่ไม่มีโครงสร้าง ทำให้ต้องพัฒนาวิธีการทางสถิติใหม่ๆ ให้เท่าทันกับการวิเคราะห์ข้อมูลที่ไม่มีโครงสร้าง สถิติกราฟิกและการสร้างภาพนิทัศน์ (Statistical Graphic and Data Visualization) กลับมีความสำคัญมาก

ยิ่งขึ้น โดยเฉพาะในปัจจุบันข้อมูลมีความซับซ้อนและยุ่งยากมากขึ้นต้องการสื่อสารให้คนทั่วไปเข้าใจได้ง่ายที่สุด ต้องเป็นนักเล่าเรื่อง (Story teller) ที่ดี

3. ข้อมูลมีความหลากหลาย Data Scientist พัฒนาวิธีการทางสถิติสำหรับการวิเคราะห์ข้อมูลข้อความ ข้อมูลเสียง ข้อมูลรูปภาพ ข้อมูลวีดีโอ ข้อมูล 3D animation ข้อมูลจาก social media ข้อมูลรูปแบบหลากหลายเหล่านี้ต้องพัฒนาวิธีการทางสถิติในการวิเคราะห์ให้ก้าวตามได้ทัน

4. ข้อมูลขนาดใหญ่มีความหลากหลาย Big data นั้นทำให้คนคาดหวังว่าจะนำข้อมูลไปสร้าง Competitive Intelligence ดังนั้นการสร้างแบบจำลองพยากรณ์ (Predictive Modeling) จากข้อมูลหลากหลายประเภทที่ไม่เคยทำมาก่อนจะยิ่งทวีความสำคัญ เช่น ต้องการพยากรณ์ว่าคนเข้าเฟซบุ๊กคนไหนน่าจะซื้อสินค้าอะไรจากข้อความ ภาพ เสียง วิดีโอ ที่เขาเข้าไปดูหรือที่เข้าโพสต์ซึ่งแบบจำลองทางสถิติแบบเดิม ๆ ไม่สามารถทำหน้าที่ดังกล่าวได้ดีพอ

5. การวิเคราะห์ข้อมูลขนาดใหญ่ มักมีเป้าหมาย หรือวัตถุประสงค์ไปเชื่อมโยงกับเนื้อหาในสาขาใดสาขาหนึ่งชัดเจน เช่น ชีวสารสนเทศศาสตร์ (Bioinformatics) การวิเคราะห์ธุรกิจ (Business Analytics) แพทย์สารสนเทศศาสตร์ (Medical Informatics) เป็นต้น นักสถิติจึงไม่สามารถมีเพียงความรู้ทางสถิติเพียงอย่างเดียวได้อีกต่อไป ไม่เพียงพอในการทำงาน

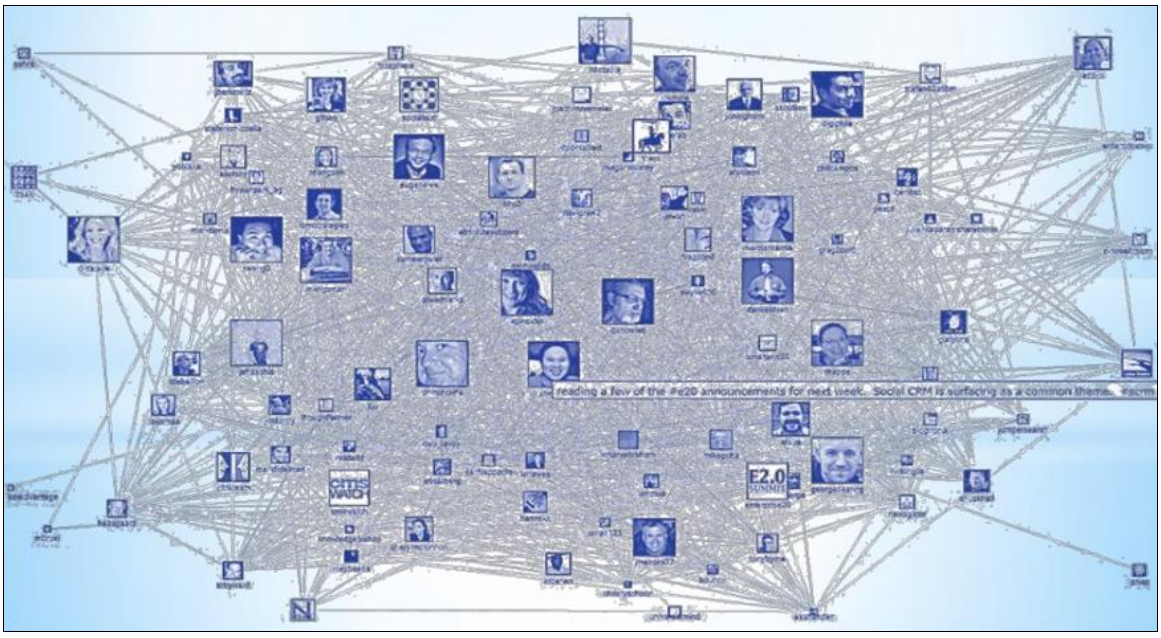
Section 2: Web and social networks

2.1 วัตถุประสงค์การเรียนรู้

1) ทราบความหมายของสื่อสังคมออนไลน์
2) ทราบประเภทของสื่อสังคมออนไลน์
3) เข้าใจอิทธิพลของสื่อสังคมออนไลน์ต่อการติดต่อสื่อสารในสังคมและผู้บริโภคและบทบาทต่อการวิเคราะห์ข้อมูลขนาดใหญ่

2.2 สื่อสังคมออนไลน์และเครือข่ายสังคมออนไลน์ (Social media and social network) คืออะไร

การวิเคราะห์ข้อมูลขนาดใหญ่ด้วยการใช้ข้อมูลจากสังคมออนไลน์ (Big Data Analytics using Social Media Data) นั้นถือว่าเป็นเรื่องที่สำคัญ เนื่องจากการวิเคราะห์เครือข่ายสังคม (social network analysis) ในปัจจุบันได้มีบทบาทต่อองค์กรต่างๆ อย่างมากมาย โดยข้อมูลที่เป็นข้อมูลจากสังคมออนไลน์อาจประกอบไปด้วย ข้อมูลจากการโทรศัพท์ การใช้ Facebook Twitter Google+ อีเมล และช่องทางอื่น ๆ อย่างที่ทราบกัน ข้อมูลเหล่านี้ถูกเรียกว่าข้อมูลเชิงดิจิทัล ซึ่งสามารถนำมาประมวลผลได้ ซึ่งข้อมูลเหล่านี้มีค่ามากมายมหาศาลในอนาคตดังรูปที่ 2.1 ข้อมูลสังคมออนไลน์ที่มีจำนวนมหาศาล



รูปที่ 2.1 ข้อมูลสังคมออนไลน์ที่มีจำนวนมหาศาล

(Smith, 2016)

สื่อสังคมออนไลน์ (Social media) คือ สื่อที่มีระบบการสื่อสารที่ใช้อินเทอร์เน็ตเป็นหลักในการติดต่อกัน ผู้ใช้สามารถสร้างเนื้อหาที่เกี่ยวข้องกับตนเองหรือที่ตนเองต้องการเพื่อแบ่งปันกับเพื่อนหรือต้องการให้สาธารณะรับรู้ก็สามารถทำได้ด้วยข้อมูลประเภทต่างๆ ทั้งตัวเลข ตัวอักษร รูปภาพเสียง วิดีโอ หรือชุดข้อมูล ผู้ใช้งานมีการเปลี่ยนสถานะเป็นได้ทั้งผู้ส่งสารและผู้รับสาร รวมถึงสามารถดึงข้อมูลที่ต้องการจากผู้ใช้งานคนอื่นได้ และการใช้งานในเรื่องการติดต่อสื่อสารกับผู้อื่นสามารถทำได้มากกว่าสื่อสารกับเพื่อนของตนเอง แต่ผู้ใช้งานสามารถทำความรู้จักกับผู้คนที่อยู่ในระบบได้ โดยสื่อสังคมออนไลน์ประกอบไปด้วยช่องทางการสื่อสารออนไลน์ต่าง ๆ ที่มีคุณลักษณะเหมือนกัน 5 อย่าง ได้แก่

- การมีส่วนร่วม (Participation) ผู้คนสามารถสร้างเนื้อหาของตนเองขึ้นมาได้และมีการให้ผลตอบรับได้ทันที
- การเปิดรับ (Openness) สื่อสังคมออนไลน์ส่วนใหญ่อนุญาตให้ผู้ใช้งานสามารถโพสต์เพื่อทำการกระจายเนื้อหาของตนเองได้
- การสนทนา (Conversation) ทำให้การสื่อสารโต้ตอบสะดวกและง่ายขึ้น
- ชุมชน (Communities) รวบรวมกลุ่มคนที่มีความสนใจร่วมกันเข้าไว้ด้วยกัน และสามารถทำได้อย่างรวดเร็ว
- การเชื่อมต่อกัน (Connectedness) มีการใช้งานที่เชื่อมต่อเนื้อหาต่าง ๆ เข้าด้วยกันอย่างมากมาย

2.3 ประเภทของสื่อสังคมออนไลน์

สื่อสังคมออนไลน์แต่ละประเภทมีคุณลักษณะเด่นที่แตกต่างกัน ทั้งในเรื่องของรูปแบบการสื่อสาร การสร้างปฏิสัมพันธ์ สื่อสังคมออนไลน์ประเภทต่าง ๆ สามารถนำมาใช้งานร่วมกันได้เพื่อทำเข้าถึงผู้ใช้ให้มากที่สุด โดยประเภทของสื่อสังคมออนไลน์มีดังต่อไปนี้

- Social networks เว็บไซต์ส่วนบุคคลที่ผู้คนจะมีการสร้างเนื้อหาหรือเรื่องราวของตนเองและทำการส่งต่อและสื่อสารไปยังเพื่อนในเครือข่ายของตน เช่น Facebook, Myspace, Bebo
- Blogs บันทึกประจำวันออนไลน์ที่คนสามารถโพสต์เรื่องราวของตนเองลงไปและให้คนอื่นๆเข้ามาแสดงความคิดเห็นได้
- Wikis หน้าเว็บไซต์ที่คนมาร่วมกันสร้างเนื้อหาและเรื่องราวและอนุญาตให้คนอื่นๆสามารถช่วยกันแก้ไขเนื้อหาได้ เช่น Wikipedia
- Podcasts ไฟล์ภาพหรือเสียงที่ถูกสร้างขึ้นและกระจายต่อไปยังคนที่ติดตามช่องทางที่เจ้าของไฟล์เหล่านี้เปิดอยู่
- Forums บอร์ดสนทนาออนไลน์ที่เปิดเพื่อให้เกิดการถกเถียงกันเกี่ยวกับหัวข้อที่น่าสนใจต่าง ๆ
- Content communities สถานที่ที่คนจัดตั้งขึ้นกันเองเพื่อสร้างเนื้อหาหรือเรื่องราวในสิ่งที่น่าสนใจเฉพาะเรื่องและให้คนอื่นๆมาแสดงความคิดเห็นได้ เช่น Youtube, Flickr

- Microblogs สถานที่ที่ผู้คนจะแบ่งปันข้อมูลในปริมาณน้อยให้กันผ่านโพสต์ของตน เช่นTwitter
- Aggregators เครื่องมือที่ทำการรวบรวมเนื้อหาต่าง ๆ จากข่าวหรือบล็อกมารวมกันให้อยู่ในที่นี่ที่เดียว โดยที่เนื้อหาที่ทำการรวบรวมมานั้นจะถูกจัดอันดับโดยความนิยมของผู้ใช้งาน รวมไปถึงผู้ใช้งานทุกคนสามารถร่วมแสดงความคิดเห็นถึงเนื้อหานั้น ๆ ได้ เช่น Reddit, Popurls
- Social bookmarking เครื่องมือที่ผู้คนใช้ในการแบ่งปันและให้คะแนนกับข้อมูลที่ตนเจอบนโลกออนไลน์และมีความน่าสนใจ เช่น Delicious

จากงานวิจัยของ Whiting และ Williams (2013) พบว่าผู้คนใช้สื่อสังคมออนไลน์ในการสร้างปฏิสัมพันธ์สูงถึง 88% หาข้อมูลข่าวสาร 80% และทำการแชร์ข้อมูลเหล่านั้นอีก 40% ด้วยความสามารถในการสร้างปฏิสัมพันธ์และส่งต่อข้อมูลข่าวสารที่รวดเร็วนี้เองทำให้เกิดเป็นเครือข่ายสังคมออนไลน์ (Social network) ขึ้น

2.4 อิทธิพลของสื่อสังคมออนไลน์ต่อการติดต่อสื่อสารในสังคมและผู้บริโภค

ผู้ใช้ที่พยายามนำเรื่องในพื้นที่ส่วนตัวของตนเองออกสู่พื้นที่สาธารณะ กลุ่มผู้ใช้งานประเภทนี้คือคนที่ใช้สื่อสังคมออนไลน์เพื่อการติดต่อสื่อสาร ทำความรู้จักกับผู้อื่นโดยการบอกนิสัยความชอบ หรือรสนิยมของตนเองในสื่อสังคมออนไลน์ เช่น การอัปโหลดวิดีโอลงบนยูทูบ จากนั้นจึงทำการติดป้ายค้นหา (Tags) ที่เป็นที่นิยมเพื่อให้ผู้ใช้คนอื่นสามารถค้นหาวิดีโอของตนได้ง่ายผ่านเว็บไซต์กูเกิล ผู้ใช้ที่ใช้สื่อสังคมออนไลน์เพื่อติดต่อสื่อสารกับเพื่อนหรือสังคมของตนโดยเลือกที่จะปกปิดข้อมูลส่วนตัว กลุ่มผู้ใช้งานประเภทนี้จะมีการติดต่อสื่อสาร และสร้างสังคมในสื่อสังคมออนไลน์เช่นกัน แต่เลือกที่จะปกปิดข้อมูลส่วนตัวบางอย่างไม่ให้เผยแพร่ไปสู่พื้นที่สาธารณะกลุ่มผู้ใช้งานทั้ง 2 มีการใช้งานสื่อเพื่อเป็นส่วนหนึ่งในเครือข่ายสังคมออนไลน์ทั้งรูปแบบของพื้นที่ส่วนตัวและพื้นที่สาธารณะ แต่ความแตกต่างของทั้ง 2 กลุ่มคือ วิธีการที่ใช้เลือกข้อมูลว่าสิ่งใดที่อยากบอกให้คนอื่นรู้ในพื้นที่สาธารณะและข้อมูลใดที่จะเก็บไว้ในพื้นที่ส่วนตัวไม่ให้ออกสู่สาธารณะ อินเทอร์เน็ตนั้นช่วยให้กลุ่มผู้ส่งสาร สามารถเลือกส่งสารในรูปแบบสาธารณะของตนไปยังพื้นที่ส่วนตัวของผู้รับสารได้เฉพาะเจาะจงมากยิ่งขึ้น เช่น การโฟกัสกลุ่มเป้าหมายทางการตลาด และผู้ส่งสารจากพื้นที่ส่วนตัวสามารถจำกัดกรอบความเป็นสาธารณะของสารที่ส่งออกไปได้ด้วยเช่นกัน

Section 3: Clustering big data Clustering social networks Apply hierarchical clustering Apply k-means clustering

3.1 วัตถุประสงค์การเรียนรู้

1) ทราบความหมายของการแบ่งกลุ่มข้อมูล
2) เข้าใจระเบียบวิธีการแบ่งกลุ่มแบบ K-means Clustering
3) เข้าใจระเบียบวิธีการแบ่งกลุ่มแบบ Hierarchical Clustering

3.2 การแบ่งกลุ่มข้อมูล

การแบ่งกลุ่มข้อมูล (clustering) เป็นการวิเคราะห์ข้อมูลที่นิยมใช้ในการเรียนรู้ของเครื่อง และการทำเหมืองข้อมูลโดยจะจัดกลุ่มของข้อมูลสำรวจ (ซึ่งมักจะอยู่ในรูปเวกเตอร์) ให้เป็นเซ็ตย่อย (เรียกว่า กลุ่ม หรือ cluster) โดยที่ข้อมูลที่มีคุณลักษณะเดียวกันจะถูกจัดกลุ่มรวมไว้ในกลุ่มเดียวกัน การแบ่งกลุ่มข้อมูลจัดเป็นวิธีการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) และเป็นวิธีที่ใช้กันทั่วไปในการวิเคราะห์ข้อมูลทางสถิติ ซึ่งขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มข้อมูลจะอาศัยความคล้าย(similarity) หรือความใกล้ชิด(proximity) โดยการวัดระยะห่างระหว่างเวกเตอร์ของข้อมูลด้วยการวัดระยะแบบต่างๆ เช่นระยะห่างแบบยุคลิด(Euclidean distance), การวัดระยะแบบแมนฮัตตัน (Manhattan (City-block) distance), การวัดระยะแบบเชบิเชฟ (Chebychev distance) สำหรับระเบียบวิธีการแบ่งกลุ่มที่นิยมใช้ได้แก่ k-means clustering, hierarchical clustering, self-organizing map (som)

3.3 ระเบียบวิธีการแบ่งกลุ่มแบบ K-means

K-means หรือเรียกอีกอย่างหนึ่งว่า การวิเคราะห์กลุ่มแบบไม่เป็น ขั้นตอน (Nonhierarchical Cluster Analysis) หรือ การแบ่งส่วน (Partitioning) เป็นรูปแบบการเรียนรู้แบบไม่มีผู้สอนที่ง่ายที่สุด เป็นการแก้ปัญหาการจัดกลุ่มที่รู้จักกันทั่วไป โดยระเบียบวิธีแบบ K-Means จะตัดแบ่ง (Partition) ข้อมูลออกเป็น K กลุ่ม และแทนค่าของแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลาง (centroid) ของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยระเบียบวิธีการจัดกลุ่มแบบ K-means มีขั้นตอนดังนี้

1) สุ่มค่าเริ่มต้น และกำหนดจุดศูนย์กลางเริ่มต้น k ตำแหน่ง เรียกว่า cluster centers หรือ centroid
2) ทำการจัดกลุ่มข้อมูลทั้งหมดโดยการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางที่กำหนดใน (1) โดยกำหนดให้ข้อมูลใดๆ ถูกจัดอยู่ในกลุ่มที่มีจุดศูนย์กลางใกล้ที่สุด
3) หาค่าเฉลี่ย (Mean) ของแต่ละกลุ่มและกำหนดให้เป็นค่าจุดศูนย์กลางใหม่
4) ดำเนินการ (2) ซ้ำจนกระทั่งค่าเฉลี่ยหรือจุดศูนย์กลางในแต่ละกลุ่มจะคงที่

สำหรับสมการการวัดระยะห่างระหว่างจุดข้อมูล $\vec{p}$ และ $\vec{q}$ แบบต่าง ๆ ใน (2) แสดงดังสมการ และรูปที่ 3.1 แสดงแผนภาพระยะห่างแบบต่างๆ

การวัดระยะห่างแบบยุคลิด

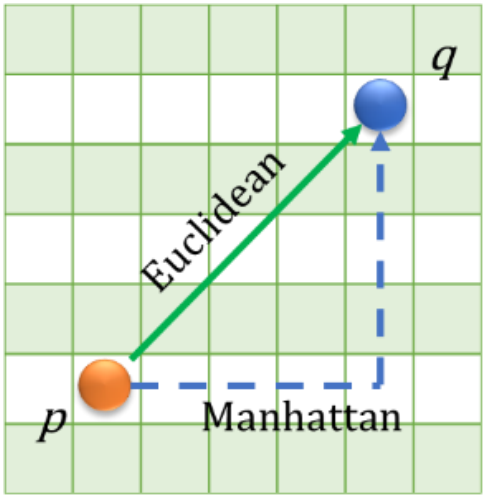
- กรณีระนาบ 2 มิติ (x, y)
- $d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$
- กรณีทั่วไป hyperplane
- $d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$

การวัดระยะห่างแบบแมนฮัตตัน

- กรณีระนาบ 2 มิติ (x, y)
- $d(p,q) = |p_x - q_x| + |p_y - q_y|$

การวัดระยะห่างแบบเชบิเชฟ

- กรณีระนาบ 2 มิติ (x, y)
- $d(p,q) = \max(|p_x - q_x|, |p_y - q_y|)$



รูปที่ 3.1 แผนภาพระยะห่างแบบต่าง ๆ

3.4 ระเบียบวิธีการแบ่งกลุ่มแบบ Hierarchical Clustering

รูปแบบนี้เป็นการจัดกลุ่มแบบมีขั้นตอน เป็นเทคนิคที่นิยมใช้ในการจัดกลุ่มเทคนิคหนึ่งซึ่งมีข้อจำกัดคือ จำนวนกลุ่มตัวอย่างที่ต้องการจัดและจำนวนตัวแปรต้องไม่มากนัก (ไม่ควรเกิน 200 กลุ่ม) ข้อดีของวิธีการนี้คือ เป็นการเรียนรู้แบบไม่มีผู้สอน ดังนั้นจึงไม่จำเป็นต้องทราบจำนวนกลุ่มที่มีในข้อมูล และไม่จำเป็นต้องทราบว่าตัวแปรใด หรือกรณีใดอยู่ในกลุ่มใด ทั้งนี้ชนิดของข้อมูลหรือตัวแปรที่สามารถใช้เทคนิค Hierarchical Cluster ได้มี 3 ประเภท คือ ข้อมูลเป็นสเกลอันตรภาค (Interval scale) หรือสเกลอัตราส่วน (Ratio scale) ซึ่งเป็นข้อมูลเชิงปริมาณ , ข้อมูลที่อยู่ในรูปความถี่ (Count Data), ข้อมูลอยู่ในรูปฐานสองนั่นคือ มีได้ 2 ค่า คือ 0 กับ 1 โดยเทคนิค Hierarchical Cluster แบ่งเป็น 2 เทคนิคย่อยคือ (Chatdanai, 2017)

- Agglomerative Hierarchical Cluster Analysis
- Divisive Hierarchical Cluster Analysis

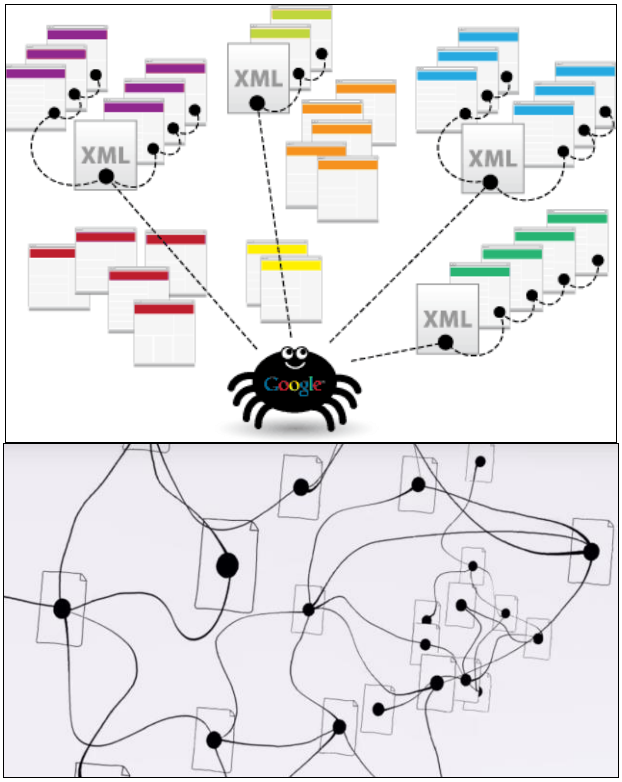
Section 4: Google web search

4.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจกระบวนการทำงานของ Google Web Search

4.2 กระบวนการทำงานของ

ปกติการค้นหาข้อมูลของ Google นั้นใช้หลักการพื้นฐานในลักษณะของการค้นคืนสารสนเทศ (Information Retrieval) อย่างไรก็ตามในระบบการค้นหาของ Google นั้นมีเทคนิคมากมายเพื่อทำให้การค้นหานั้นได้รับคำตอบที่ตรงเป้าหมายและรวดเร็วที่สุด ขั้นตอนโดยภาพรวมประกอบด้วย



รูปที่ 4.1 หลักการของโปรแกรม Spider ในระบบการค้นหาของ Google

(Google, 2019)

1) การค้นหาและรวบรวมข้อมูล: ขั้นตอนแรกสุด Google ใช้โปรแกรมที่ถูกเรียกว่า “Spider” วิ่งไปตามเว็บไซต์และวิ่งไปตามจุดเชื่อมโยง (links) ต่าง ๆ เพื่อทำการเก็บข้อมูลที่เกี่ยวข้องให้ได้มากที่สุด ซึ่งข้อมูลที่ถูกบรรจุไว้ใน Metadata ของเว็บไซต์นั้นเป็นส่วนสำคัญที่สุดที่ตัว Spider สามารถค้นหาข้อมูลได้ หลักการทำงานดังกล่าวแสดงได้ดังรูปที่ 4.1

2) การจัดระเบียบข้อมูลด้วยการจัดทำดัชนี: เมื่อโปรแกรมรวบรวมข้อมูลที่พบบนหน้าเว็บ ระบบของ Google จะทำการจัดทำดัชนี (index) เพื่อการเข้าถึงที่รวดเร็วขึ้น คล้าย ๆ กับการจัดเก็บข้อมูลดัชนีที่อยู่ในท้ายหนังสือ เมื่อมีการพบหน้าเว็บใหม่ ก็จะมีการจัดทำดัชนีเพิ่มเติมไปเรื่อย ๆ ดังรูปที่ 4.2

รูปที่ 4.2 การจัดระเบียบข้อมูลด้วยการจัดทำดัชนี

(Google, 2019)

3) การค้นหาเพจเป้าหมายและผลคะแนน: หากเราต้องการค้นหาคำว่า “เสือชีตาห์” แน่นอนว่าข้อมูลที่ได้นั้นอาจมีมากเป็น 1,000,000 เพจที่เกี่ยวข้อง ดังนั้นคำถามที่น่าสนใจคือ ทำอย่างไรสำหรับการแสดงผลข้อมูลที่เกี่ยวข้อง โดยให้เรียงลำดับเพจให้ตรงเป้าหมายที่สุด ซึ่ง Google นั้นมีขั้นตอนวิธีเฉพาะ โดยใช้วิธีการคำนวณน้ำหนักของคำที่เกี่ยวข้อง เช่น a, an, the (article) in, และ on (preposition) เป็นต้น คำพวกนี้ส่วนมากความสำคัญจะน้อยกว่าคำเฉพาะ ดังนั้นน้ำหนักในการนำมาคำนวณคะแนนจะน้อยกว่า เป็นต้น

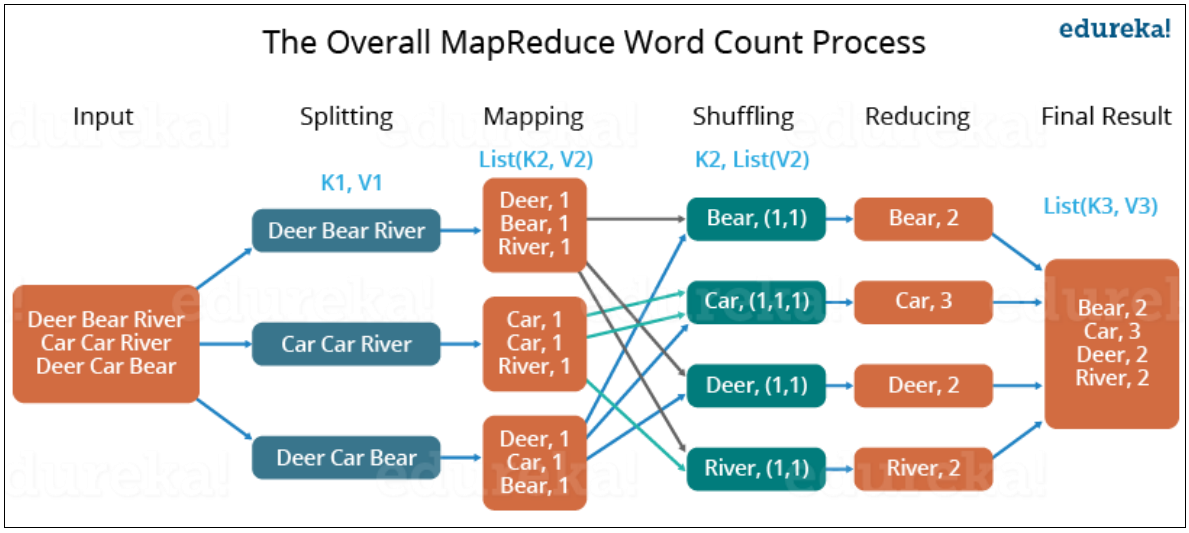
Section 5: Parallel and distributed computing using MapReduce

5.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจกระบวนการทำงานแบบขนานและกระจายตัวด้วย Map/Reduce

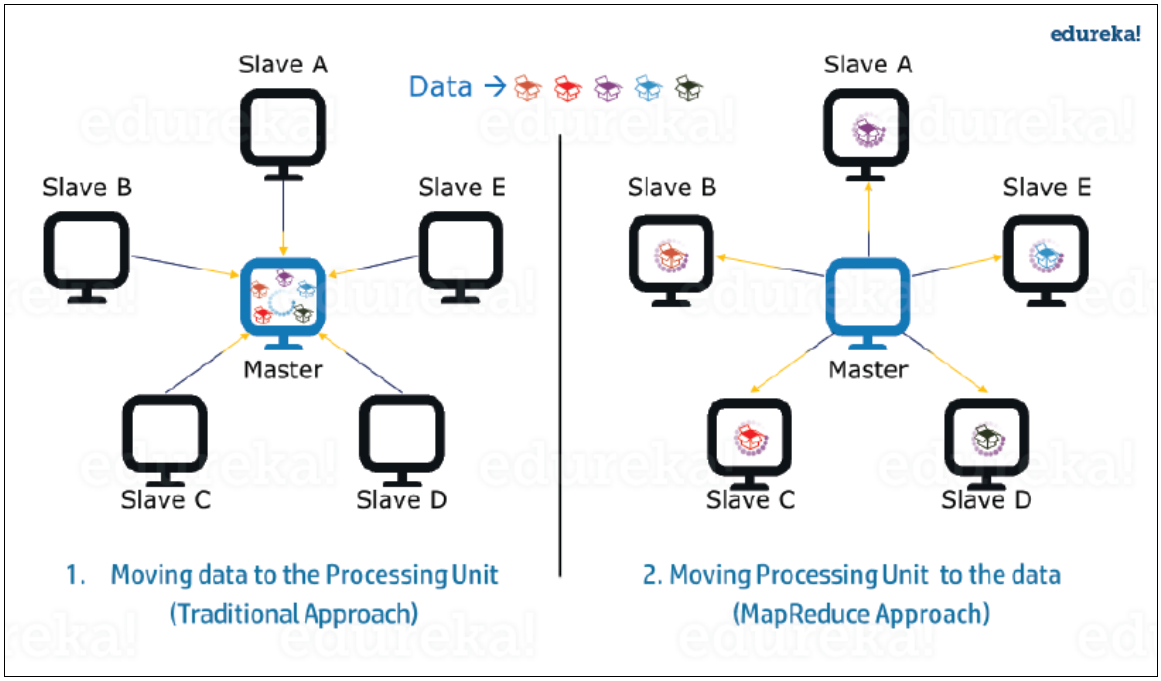
5.2 กระบวนการทำงานแบบขนานและกระจายตัวด้วย Map/Reduce

Map/Reduce จะเป็นส่วนประมวลผลข้อมูล ที่นักพัฒนาสามารถเขียน โปรแกรมโดยใช้ภาษาจาวามาวิเคราะห์ข้อมูลในรูปแบบของฟังก์ชันการ Map และ Reduce ได้โดยระบบก็จะกระจาย Task ไปรันแบบ Parallel บนเครื่องหลายๆ เครื่องดังรูปที่ 5.1 และ รูปที่ 5.2



รูปที่ 5.1 ภาพรวมของ MapReduce Word Count Process

(Adam & Josh, 2017)



รูปที่ 5.2 ตัวอย่างของ MapReduce

(Adam & Josh, 2017)

ข้อมูลที่เก็บอยู่ใน HDFS จะไม่ใช่รูปแบบ Table อย่างที่เก็บในฐานข้อมูล RDBMS จะเหมาะกับการเก็บข้อมูลขนาดใหญ่มากที่ไม่ต้องมีการเปลี่ยนแปลง และไม่สามารถอ่านหรือเขียนข้อมูลแบบ Random Access ได้ส่วนการประมวลผลแบบ Map/Reduce ก็ไม่ใช่ real time Online แบบ SQL ของ RDBMS แต่จะเป็นแบบ Batch Online ใช้เวลาพอสมควรขึ้นอยู่กับขนาดข้อมูล สถาปัตยกรรมฮาร์ดแวร์ของระบบ Hadoop จะประกอบด้วยเครื่อง Server จำนวนมาก โดยจะมีเครื่องหนึ่งทำหน้าที่เป็น Master และจะมีเครื่องลูกอีกจำนวนมากทำหน้าที่เป็น Slave โดยปกติ Hadoop จะกำหนดให้ข้อมูลที่เก็บในเครื่อง Slave มีการเก็บข้อมูลซ้ำกันสามแห่ง ดังนั้นเครื่อง Slave ควรจะมีอย่างน้อยสามเครื่อง ส่วนเครื่อง Master ก็จะทำหน้าที่หลัก ในการระบุตำแหน่งของข้อมูลและ Task ที่กระจายในการประมวลผลของ Map/Reduce ดังนั้นเครื่อง Master จึงมีความสำคัญอย่างมาก และต้องมีเครื่อง Secondary Master ในการที่จะสำรองไว้ในกรณีเครื่อง Master ตายไป ดังนั้นระบบ Hadoop โดยทั่วไปจะเริ่มต้นที่เครื่อง Server 5 เครื่อง สำหรับ Master หนึ่งเครื่อง Secondary Master หนึ่งเครื่อง และ Slave สามเครื่อง โดยหากต้องการเก็บข้อมูลมากขึ้นหรือต้องการประมวลผลข้อมูลให้เร็วขึ้นก็ต้องเพิ่มจำนวนเครื่อง Slave ให้มากขึ้น

Section 6: Computing similar documents in big data

6.1 วัตถุประสงค์การเรียนรู้

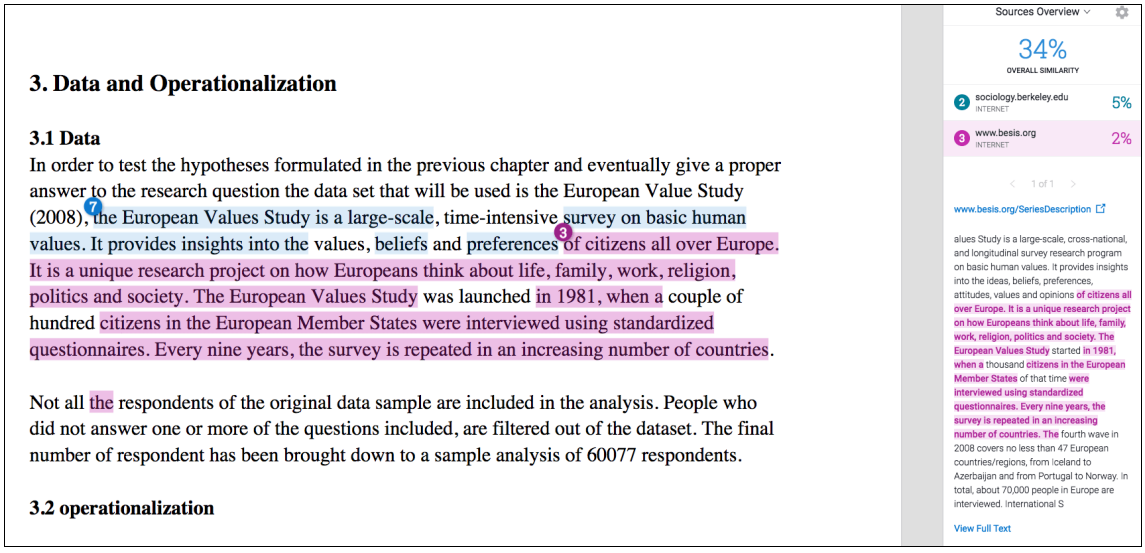
1) เข้าใจการเปรียบเทียบเอกสารที่เหมือนกันด้วยเทคโนโลยีของ Big Data

6.2 กรณีศึกษาการเปรียบเทียบเอกสารที่เหมือนกันด้วยเทคโนโลยีของ Big Data

การเปรียบเทียบเอกสารที่เหมือนกันด้วยเทคโนโลยีของ Big Data ถือว่าเป็นสิ่งที่ท้าทาย เนื่องจากการมีอยู่ของเอกสารที่อยู่ในระบบอินเทอร์เน็ตนั้นมีอยู่เป็นจำนวนมาก ดังนั้นการค้นหาเอกสารที่มีความใกล้เคียงกันจะมีประโยชน์อย่างมาก เช่น การจัดกลุ่มเอกสารที่มีความใกล้เคียงกัน การเรียกดูเอกสารที่มีความซ้ำซ้อนกันเพื่อทำลายทิ้ง เป็นต้น

ปัจจุบันมีเครื่องมือมากมายที่สามารถตรวจสอบความเหมือนกันของเอกสารได้ เช่น โปรแกรมTurnitin เป็นต้น โดยใช้หลักการพื้นฐานในลักษณะของการค้นคืนสารสนเทศ

กรณีศึกษาที่เห็นได้ชัดเจนที่สุดคือการตรวจสอบการคัดลอกผลงาน (Plagiarism) เนื่องจากการคัดลอกผลงานในวงการวิชาการถือว่าเป็นสิ่งที่ยอมรับไม่ได้ รูปที่ 6.1 แสดงตัวอย่างหน้าจอโปรแกรมสำหรับตรวจสอบการคัดลอกผลงาน ซึ่งจะปรากฏเอกสารที่มีความเหมือนกันด้านขวา นอกจากนี้ยังบอกถึงร้อยละของความเหมือนกันด้วย โดยการวิเคราะห์ความเหมือนของเอกสารจะใช้เทคนิค Map-Reduced คู่กับอัลกอริทึม SCAM (Standard. Copy Analysis Mechanism)



3. Data and Operationalization

3.1 Data

In order to test the hypotheses formulated in the previous chapter and eventually give a proper answer to the research question the data set that will be used is the European Value Study (2008), the European Values Study is a large-scale, time-intensive survey on basic human values. It provides insights into the values, beliefs and preferences of citizens all over Europe. It is a unique research project on how Europeans think about life, family, work, religion, politics and society. The European Values Study was launched in 1981, when a couple of hundred citizens in the European Member States were interviewed using standardized questionnaires. Every nine years, the survey is repeated in an increasing number of countries.

Not all the respondents of the original data sample are included in the analysis. People who did not answer one or more of the questions included, are filtered out of the dataset. The final number of respondent has been brought down to a sample analysis of 60077 respondents.

3.2 operationalization

Sources Overview

34%

OVERALL SIMILARITY

2 sociology.berkeley.edu INTERNET 5%

3 www.besis.org INTERNET 2%

1 of 1

www.besis.org/SeriesDescription

alues Study is a large-scale, cross-national, and longitudinal survey research program on basic human values. It provides insights into the ideas, beliefs, preferences, attitudes, values and opinions of citizens all over Europe. It is a unique research project on how Europeans think about life, family, work, religion, politics and society. The European Values Study started in 1981, when a thousand citizens in the European Member States of that time were interviewed using standardized questionnaires. Every nine years, the survey is repeated in an increasing number of countries. The fourth wave in 2008 covers no less than 47 European countries/regions, from Iceland to Azerbaijan and from Portugal to Norway. In total, about 70,000 people in Europe are interviewed. International S

View Full Text

รูปที่ 6.1 ตัวอย่างหน้าจอโปรแกรมสำหรับตรวจสอบการคัดลอกผลงาน

(Petersen, 2016)

Section 7: Products frequently bought together in stores

7.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจการหาความสัมพันธ์ของข้อมูลด้วยเทคโนโลยีการวิเคราะห์ข้อมูลขนาดใหญ่

7.2 กรณีศึกษาการหาความสัมพันธ์ของข้อมูลด้วยเทคโนโลยีการวิเคราะห์ข้อมูลขนาดใหญ่

กรณีศึกษาในห้างสรรพสินค้า สามารถนำเอาเทคโนโลยีการวิเคราะห์ข้อมูลขนาดใหญ่มาหาความสัมพันธ์ของข้อมูล ซึ่งการค้นหาความสัมพันธ์และรูปแบบ (Pattern) ทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก เทคนิคดังกล่าวเป็นเทคนิคของการทำเหมืองข้อมูล หรือ “Data Mining” ซึ่งจะทำการสำรวจและวิเคราะห์อย่างอัตโนมัติหรือกึ่งอัตโนมัติ ในปริมาณข้อมูลจำนวนมากให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ (Rule) โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่าง ๆ ที่มีประโยชน์ในฐานข้อมูล

กรณีศึกษาที่ชัดเจนที่สุดคือ “ห้างสรรพสินค้าแห่งหนึ่ง ค้นพบพฤติกรรมของผู้บริโภค ที่พ่อบ้านมักซื้อเบียร์และผ้าอ้อมในวันศุกร์ตอนเย็น” ดังนั้นเป็นสัญญาณให้เจ้าของกิจการควรจะเตรียมสินค้าไว้เพื่อจำหน่าย หรือจัดวางสินค้าไว้ในบริเวณเดียวกัน เป็นต้น

อย่างไรก็ตามในกรณีนี้ ข้อมูลที่นำมาวิเคราะห์ควรเป็นข้อมูลขนาดใหญ่จริง ๆ และข้อมูลตั้งต้นต้องมีความถูกต้องด้วย ซึ่งบางครั้งการใช้ข้อมูลมักจะพบว่ามีความสัมพันธ์ของการซื้อสินค้า 2 อย่างเสมอ เมื่อจำนวนความหลากหลายของสินค้ามากขึ้น แต่ไม่ได้หมายความว่าจะต้องให้ห้างสรรพสินค้าเก็บสินค้าในคลังมากขึ้น เพราะข้อมูลที่ได้อาจเกิดความคลาดเคลื่อน เพราะฉะนั้นจะต้องทำการตรวจสอบความถูกต้องของข้อมูลและวิเคราะห์ความถูกต้องอีกครั้ง

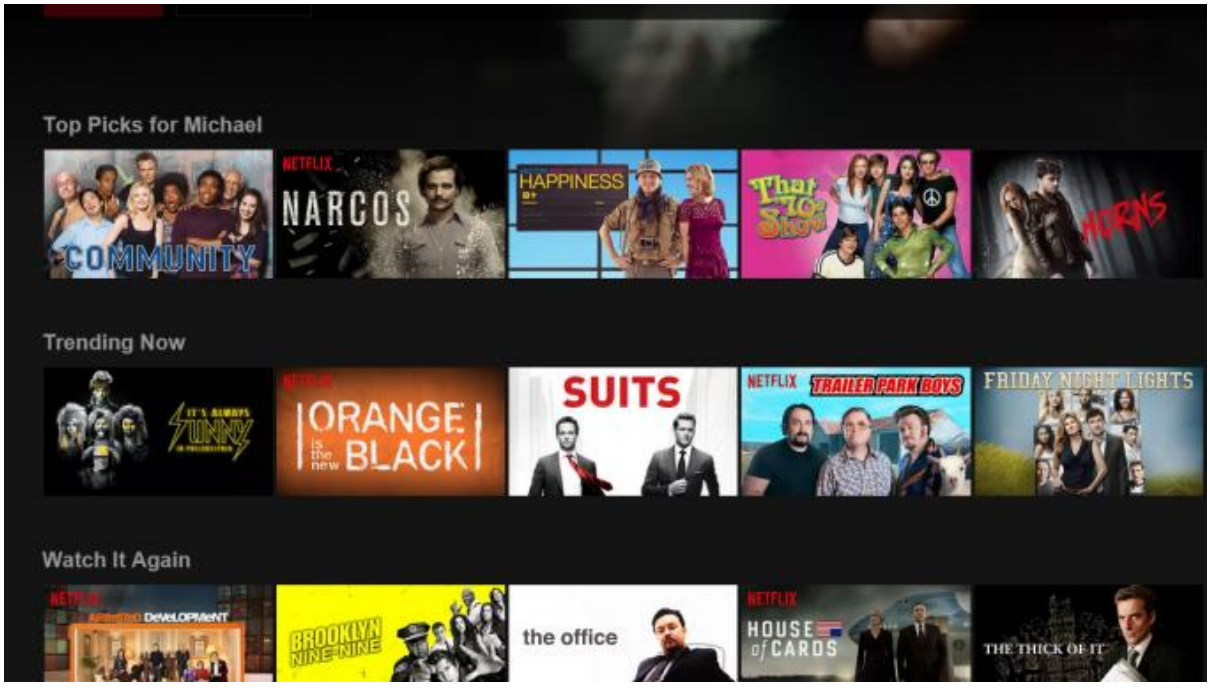
Section 8: Movie and music recommendations

8.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจการแนะนำข้อมูลภาพยนต์และเพลงด้วยเทคโนโลยีการวิเคราะห์ข้อมูลขนาดใหญ่

8.2 กรณีศึกษาการแนะนำข้อมูลภาพยนต์และเพลงด้วยเทคโนโลยีการวิเคราะห์ข้อมูลขนาดใหญ่

คำถามที่น่าสนใจในยุคปัจจุบันคือทำไมบริษัท Netflix จึงสามารถสร้างรายชื่อภาพยนตร์แนะนำที่ตรงใจลูกค้าแต่ละคน (personalize) ได้ คำตอบที่สมเหตุผลที่สุดคือการที่บริษัท Netflix ได้นำ Big Data มาประมวลผลเพื่อเข้าถึงลูกค้าให้ได้มากที่สุด ปัจจุบันผู้ผลิตสื่อและเนื้อหาต่าง ๆ สามารถนำเสนอข้อมูลที่ตรงใจผู้อ่านได้เพียงแค่กดดูข้อมูลจำนวนมหาศาล จากนั้นผู้ผลิตสื่อสามารถนำข้อมูลเหล่านี้มาวิเคราะห์ว่าผู้คนกำลังสนใจที่จะอ่านเรื่องใดมากที่สุด ดังนั้นผู้ผลิตสื่อจึงได้ข้อมูลเฉพาะของลูกค้า เพื่อนำเสนอเนื้อหาหรือเขียนเรื่องราวให้ตรงจุดที่สุดเท่าที่จะเป็นไปได้ นอกจากนี้นักการตลาดสายดิจิตอลยังสามารถนำข้อมูลจากเว็บไซต์มาใช้ประโยชน์ได้ เช่น ข้อมูลที่อยู่ อายุ เพศ ของผู้ที่เข้าชมเว็บไซต์มาพัฒนาเพื่อให้เข้าถึงกลุ่มเป้าหมายได้มากขึ้นอีก ดังที่ได้กล่าวไปข้างต้น Netflix ใช้วิธีการเก็บข้อมูลผู้ชมภาพยนตร์และประมวลผลข้อมูล จนสามารถพัฒนาเนื้อหารายการใหม่ ๆ ที่น่าจะถูกใจผู้ชม และสามารถแนะนำรายการที่น่าสนใจสำหรับแต่ละบุคคลได้ รูปที่ 8.1 แสดงตัวอย่างหน้าจอการนำเสนอภาพยนตร์ที่ตรงใช้ผู้ใช้ของ Netflix



รูปที่ 8.1 ตัวอย่างหน้าจอการนำเสนอภาพยนตร์ที่ตรงใช้ผู้ใช้ของ Netflix

(Petersen, 2016)

ในขณะเดียวกัน ทางบริษัท Spotify ผู้นำตลาดสื่อดิจิทัลเพลง ได้มีการปรับปรุงหน้าการค้นหา โดยใช้คุณสมบัติของข้อมูลลูกค้าที่เข้ามาทำกิจกรรมในระบบ ซึ่งข้อมูลนั้นมีจำนวนมากมายมหาศาล เพื่อสร้างคำแนะนำและเสริมสร้างประสบการณ์ให้สอดคล้องกับความชอบที่มีลักษณะแตกต่างกันของผู้ใช้ การทำงานดังกล่าวของบริษัท Spotify ก็มีลักษณะเช่นเดียวกับ Netflix

Section 9: Google's AdWords TM System

9.1 วัตถุประสงค์การเรียนรู้

1) ทราบความหมายของเครื่องหมายการค้าและ Google's AdWords

9.2 เครื่องหมายการค้า

เครื่องหมายการค้า (Trademark) หมายถึง ตราสินค้าหรือส่วนหนึ่งของตราสินค้า เพื่อแสดงว่าสินค้าที่ใช้เครื่องหมายของเจ้าของเครื่องหมายการค้านั้น เจ้าของมีสิทธิตามกฎหมายเพียงผู้เดียว เราไม่สามารถใช้เครื่องหมายการค้าของบุคคลอื่นและบุคคลอื่นก็ไม่สามารถใช้เครื่องหมายการค้าของเราได้ จึงต้องมีการออกแบบโลโก้ หรือสัญลักษณ์ ซึ่งอาจจะประกอบไปด้วย ชื่อ ข้อความ วลี ภาพ สัญลักษณ์ งานออกแบบ หรือหลายส่วนร่วมกัน โดยมีความหมายทางด้านทรัพย์สินทางปัญญาเป็นเครื่องหมายแสดงถึง ชื่อสินค้าเฉพาะอย่าง

9.3 Google's AdWords

Google's AdWords คือ platform ที่ให้บริการในการทำโฆษณาออนไลน์ ถูกพัฒนาและให้บริการโดย Google สำหรับให้นักโฆษณา (Advertiser) หรือเจ้าของธุรกิจที่ต้องการทำโฆษณาออนไลน์ ใช้เพื่อทำการโฆษณาออนไลน์บนเครือข่ายของ Google ได้ด้วยตนเอง เพื่อเข้าถึงกลุ่มเป้าหมายหรือกลุ่มลูกค้าของผู้โฆษณาโดยมีช่องทางที่หลากหลาย เช่น

Google Search, Google Display Network, Google partners, Gmail หรือแม้กระทั่งการโฆษณาบน Youtube ก็สามารถทำได้ผ่าน Google Adwords เช่นกัน

อาจเปรียบได้ว่า Google Adwords นั้นคือ บริการด้านสื่อโฆษณาออนไลน์ที่จะช่วยเชื่อมต่อกับกลุ่มเป้าหมายของได้ ดังนั้นการใช้บริการ Google Adwords จะสามารถส่งผลให้ Trademark ของบริษัทถูกพูดถึงและแพร่กระจายไปยังส่วนต่าง ๆ ได้ถึงกลุ่มเป้าหมายได้ดีขึ้น

Section 10: Mining rapidly arriving data streams

10.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจกระบวนการทำเหมืองข้อมูลจากข้อมูลที่เข้ามาแบบ real-time streaming

10.2 ความหมายของ Social Media Listening

กระบวนการทำเหมืองข้อมูลจากข้อมูลที่มีเข้ามาเรื่อยๆ มีการเปลี่ยนแปลงตลอดเวลานั้นถือว่ามีส่วนสำคัญในแง่ของธุรกิจ ซึ่งส่วนมากข้อมูลเหล่านี้คือการรับฟังความคิดเห็นของผู้บริโภค(ในแง่ธุรกิจ) ผ่านทางโซเชียลมีเดียต่างๆ ซึ่งสิ่งต่างๆ ที่ถูกถ่ายทอดผ่านทางโซเชียลมีเดียเหล่านั้น อาจมีการกล่าวถึงผลิตภัณฑ์หรือบริการของบริษัท สิ่งนี้จึงเป็นสิ่งที่สำคัญมากในการที่จะทราบถึงความคิดเห็นทั้งในเชิงบวกและเชิงลบที่มีต่อบริการเนื่องจากปัจจุบันเป็นยุคที่ทุกคนสามารถเขียนเนื้อหา(content) ผ่านช่องทางอินเทอร์เน็ตด้วยการโพสต์เฟซบุ๊ก(Facebook) ทวิตเตอร์(Twitter) และโซเชียลมีเดียอื่นๆ โดยแพลตฟอร์มการเฝ้าระวังจะกล่าวโดยการสาธิตในบทสุดท้าย

Social Media Listening คือการรับฟังความคิดเห็นของผู้บริโภคผ่านทางโซเชียลมีเดียต่างๆ ตามที่ได้เกริ่นไว้ก่อนหน้านี้ อีกทั้งผู้ที่เข้ามาอ่าน content เหล่านั้นยังสามารถที่จะแสดงความคิดเห็นจากประสบการณ์ ข้อเสนอแนะหรืออะไรก็ตามที่ต้องการจะสื่อออกมา จึงส่งผลให้ content ต่างๆ บนโซเชียลมีเดียมีค่ามากต่อหลายๆ บริษัทและแม้กระทั่งหน่วยงานรัฐบาล ข้อมูลที่ได้จาก social media listening นั้นสามารถนำไปต่อยอดได้ในหลายๆด้าน เช่น การบริหารจัดการกับวิกฤต (Crisis Management) จากการที่มีผู้ใช้พูดถึงสินค้าหรือบริการในแง่ที่ไม่ดี จึงทำหน่วยงานเจ้าของบริการนั้นสามารถรับมือได้ อีกทั้ง crisis management ยังส่งผลดีต่อชื่อเสียงของบริษัท(Brand Reputation) และยังทำให้เข้าใจมุมมองของผู้บริโภค (Customer Insight) ได้มากขึ้นเนื่องจากเป็นสิ่งที่ผู้บริโภคพูดออกมาเอง (hooktalk, 2017)

การทำ Social Media Listening นั้นอยู่คู่กับการบริหารธุรกิจหรือการบริหารองค์กรมาตั้งแต่ต้น เพียงแต่รูปแบบในการรับฟังความคิดเห็นหรือเสียงของลูกค้านั้นเปลี่ยนไปตามยุคสมัยและเครื่องมือที่ใช้ในการประกอบธุรกิจ เช่น ในอดีตในยุคที่ยังไม่มีโซเชียลมีเดีย หากบริษัทต้องการทราบความเห็นของลูกค้าก็อาจจะต้องมีการออกภาคสนามเพื่อทำ research ให้ได้มาซึ่งความคิดเห็นขอลูกค้า มีการร้องขอให้ลูกค้าตอบแบบสอบถาม แต่ในปัจจุบันผู้คนส่วนมากมีบัญชีโซเชียลมีเดียเป็นของตัวเอง พวกเขาเหล่านั้นจะแสดงความคิดเห็นผ่านสื่อเหล่านั้น สิ่งที่บริษัทต้องทำคือทำการติดตาม (track) และนำข้อมูลเหล่านั้นมาวิเคราะห์ แต่กระนั้นก็ตามการติดตามความคิดเห็นของลูกค้าต่อผลิตภัณฑ์หรือบริการแบบดั้งเดิมก็ยังคงถูกใช้อยู่เพราะไม่ใช่ทุกคนที่จะใช้โซเชียลมีเดียและไม่ใช่คนที่ใช้โซเชียลมีเดียทุกคนจะแสดงความคิดเห็นต่อสินค้าหรือบริการที่พวกเขาเหล่านั้นใช้ จึงต้องใช้วิธีทั้งแบบดั้งเดิมและแบบใหม่ควบคู่กันไปเพื่อให้ได้มาซึ่งข้อมูลที่เป็นเสียงต่อลูกค้าที่ครอบคลุมกลุ่มลูกค้าให้ได้มากที่สุด

10.3 ความสำคัญของ Social Media Listening

โซเชียลมีเดียนั้นเข้าถึงบุคคลแทบจะทุกเพศ ทุกวัยและเป็นสังคมเดียวในโลกใบนี้เท่านั้นที่จะสามารถพบปะพูดคุยกับคนทุกเพศทุกวัยทั่วโลกรวมถึงรับฟังความคิดเห็นหรือสิ่งที่คนเหล่านั้นคิดต่อสิ่งต่างๆ จึงไม่แปลกเลยที่ในหลายๆ บริษัทในปัจจุบันจะหันมาพัฒนาเครื่องมือและกลยุทธ์ในการใช้โซเชียลมีเดียเพื่อเหตุผลด้านการทำธุรกิจ สินค้าและบริการต่างๆ ทั่วโลกล้วนมีผู้ใช้ที่เข้าใช้โซเชียลมีเดียด้วยเช่นกัน ฉะนั้นจึงแน่ใจได้ว่าผู้บริโภคเหล่านั้นอาจจะต้องมีการแสดงความคิดเห็นต่อสินค้าที่พวกเขาใช้บนโลกออนไลน์อย่างแน่นอนไม่มากก็น้อย (Evans, 2015) และนั่นก็เป็นข้อมูลที่มีค่าต่อบริษัทอย่างยิ่งในการนำไปต่อยอดในการพัฒนาหรือปรับปรุงสินค้านั้นๆ เพื่อให้ตอบสนองความต้องการของลูกค้าให้ได้มากที่สุด

10.4 เริ่มต้นการทำ Social Media Listening

(Evans, 2015) ได้แนะแนวทางการทำ Social Media Listening อย่างมีประสิทธิภาพโดยมีขั้นตอนดังรูปที่ 10.1

รูปที่ 10.1 กระบวนการทำ Social Media Listening

(Petersen, 2016)

Section 11: Introduction to data files data format data encoding

11.1 วัตถุประสงค์การเรียนรู้

1) ทราบความหมายของ Data File, Data Format, และ Data Encoding

11.2 Data File

ข้อมูลที่ได้ขึ้นชื่อว่า Big Data นั้น จะต้องมีคุณสมบัติอย่างหนึ่งคือ ความหลากหลายของข้อมูล (variety of data) ซึ่งอาจประกอบไปด้วย

- Data file: คือไฟล์คอมพิวเตอร์ ซึ่งบรรจุเนื้อหาที่ถูกใช้ในคอมพิวเตอร์ โดยปกติแล้วจะไม่บรรจุคำสั่งของการ execute

11.3 Data Format

- Data Format: เป็นรูปแบบข้อมูลที่สามารถถูกจัดเก็บหรือสามารถสื่อสารผ่านระบบเครือข่ายคอมพิวเตอร์ได้ ประกอบด้วยรูปแบบหลายรูปแบบ ตัวอย่างเช่น
  - Data type
  - Signal
  - Recording format
  - File format
  - Content format
  - Audio format
  - Video format

11.4 Data Encoding

- Data Encoding: เป็นการเข้ารหัสข้อมูล เพื่อปกป้องข้อมูลระหว่างการถ่ายโอนซึ่งความหลากหลายของข้อมูลดังกล่าวจึงทำให้ขั้นตอนการประมวลผลข้อมูลขนาดใหญ่มีความซับซ้อนยิ่งขึ้น

Section 12: Introduction to relational database

12.1 วัตถุประสงค์การเรียนรู้

1) ทราบความหมายของฐานข้อมูลเชิงสัมพันธ์

12.2 ฐานข้อมูลเชิงสัมพันธ์

ฐานข้อมูลเชิงสัมพันธ์ นั้นหมายความว่า จะมีการจัดเก็บข้อมูลในลักษณะที่เป็นกลุ่มของข้อมูลที่มีความสัมพันธ์กัน ในฐานข้อมูลหนึ่ง ๆ สามารถที่จะมีตารางตั้งแต่ 1 ตารางเป็นต้นไป และในแต่ละตารางนั้นก็สามารถมีได้หลายคอลัมน์ (Column) หลายแถว (Row) ตัวอย่างเช่น เราต้องการเก็บข้อมูลพนักงาน ในตารางของข้อมูลพนักงานก็จะประกอบด้วยคอลัมน์ ที่อธิบายชื่อ นามสกุล ที่อยู่ เงินเดือน แผนกที่สังกัด เป็นต้น และในตารางนั้น ก็สามารถที่จะมีข้อมูลพนักงานได้มากกว่า 1 คน (Row) และตารางข้อมูลพนักงานนั้นอาจจะมีความสัมพันธ์กับตารางอื่น เช่น ตารางที่เก็บชื่อและจำนวนบุตรของพนักงาน

ฐานข้อมูลเชิงสัมพันธ์ถูกออกแบบมาเพื่อลดความซ้ำซ้อนของการเก็บข้อมูล และสามารถเรียกใช้ข้อมูลได้อย่างมีประสิทธิภาพ โดยมีหลักดังนี้

- ตารางจะต้องมีชื่อไม่ซ้ำกัน
- แต่ละฟิลด์จะบรรจุประเภทข้อมูลเพียงชนิดเดียวเท่านั้นแน่นอน
- ข้อมูลในแต่ละเรคคอร์ดจะต้องไม่ซ้ำกัน
- นอกจากนี้แต่ละตารางยังสามารถเรียกได้อีกอย่างว่ารีเลชัน(Relation) แถวแต่ละแถวภายในตารางเรียกว่าทูเพิล (Tuple) และคอลัมน์เรียกว่าแอททริบิวต์ (Attribute)

จุดเด่นของข้อมูลเชิงสัมพันธ์

- ง่ายต่อการเรียนรู้ และการนำไปใช้งาน ทำให้เห็นภาพข้อมูลชัดเจน
- ภาษาที่ใช้จัดการข้อมูลเป็นแบบซีเควล (SQL) หรือเอสคิวแอล ซึ่งมีประสิทธิภาพสูงเข้าใจง่าย
- การออกแบบระบบมีทฤษฎีรองรับ สามารถลดความซ้ำซ้อนของข้อมูลได้
- กฎที่เกี่ยวข้องกับคีย์ในฐานข้อมูลเชิงสัมพันธ์

1. กฎความบูรณภาพของเอนทิตี้ (The Entity Integrity Rule)

กฎนี้ระบุไว้ว่าแอททริบิวต์ใดที่เป็นคีย์หลักข้อมูลในแอททริบิวต์นั้นจะเป็นค่าว่าง (Null) ไม่ได้ ความหมายของการเป็นค่าว่างไม่ได้ (Not Null) หมายถึงข้อมูลของแอททริบิวต์ที่เป็นคีย์หลักจะไม่ทราบค่าที่แน่นอนหรือไม่มีค่าไม่ได้

2. กฎความบูรณภาพของการอ้างอิง (The Referential Integrity Rule)

การอ้างอิงข้อมูลระหว่างรีเลชันในฐานข้อมูลเชิงสัมพันธ์จะใช้คีย์นอกของรีเลชันหนึ่งไปตรวจสอบกับค่าของแอททริบิวต์ ที่เป็นคีย์หลักของอีกรีเลชันหนึ่ง เพื่อเรียกดูข้อมูลอื่นๆ ที่เกี่ยวข้องหรือค่าของคีย์นอกจะต้องอ้างอิงให้ตรงกับค่าของคีย์หลักได้จึงจะสามารถเชื่อมโยงข้อมูลระหว่างสองรีเลชันได้ สำหรับคีย์นอกจะมีค่าว่างได้หรือไม่ขึ้นอยู่กับกฎเกณฑ์การออกแบบฐานข้อมูล เช่น ในกรณีที่รีเลชั่นพนักงานมี Depno เป็นคีย์นอกอาจจะถูกระบุว่าต้องทราบค่า แต่ในกรณีพนักงานทดลองงานอาจยังไม่มีค่า Depno เพราะยังไม่ได้ถูกบรรจุ ในกรณีที่มีการลบ หรือแก้ไขข้อมูลของแอททริบิวต์ที่เป็นคีย์หลักซึ่งมีคีย์นอกจากอีกรีเลชันหนึ่งอ้างอิงถึง จะทำการลบหรือแก้ไขข้อมูลได้หรือไม่ขึ้นอยู่กับการออกแบบฐานข้อมูลว่าได้ระบุให้แอททริบิวต์ให้มีคุณสมบัติอย่างไร ซึ่งมีโอกาสเป็นไปได้ 4 ทางเลือก

การลบหรือแก้ไขข้อมูลแบบมีข้อจำกัด (Restrict) การลบหรือแก้ไขข้อมูลจะกระทำได้ เมื่อข้อมูลของคีย์หลักในรีเลชันหนึ่งไม่มีข้อมูลที่ถูกอ้างอิง โดยคีย์นอกของอีกรีเลชันหนึ่งเช่น รหัสแผนก Dep.no ในรีเลชัน Dep.no จะถูกแก้ไขหรือลบทิ้งต่อเมื่อไม่มีพนักงานคนใดสังกัดอยู่ในแผนกนั้น

การลบหรือแก้ไขข้อมูลแบบต่อเรียง (Cascade) การลบหรือการแก้ไขข้อมูล จะทำแบบเป็นลูกโซ่ คือ หากมีการแก้ไข หรือลบข้อมูลของคีย์หลักในรีเลชันหนึ่งระบบจะทำการลบ หรือแก้ไขข้อมูลของคีย์นอกในอีกรีเลชันหนึ่งที่อ้างอิงถึงข้อมูลของคีย์หลักที่ถูกลบให้ด้วย เช่น ในกรณีที่ยกเลิกแผนก 9 ใน Entity แผนก ข้อมูลของพนักงานที่อยู่แผนก 10 ใน Entity พนักงานจะถูกลบออกไปด้วย

การลบหรือแก้ไขข้อมูลโดยเปลี่ยนเป็นค่าว่าง (Nullify) การลบหรือแก้ไขข้อมูลจะทำได้เมื่อมีการเปลี่ยนค่าของคีย์นอกในข้อมูลที่ถูกอ้างอิงให้เป็นค่าว่างเสียก่อน เช่น พนักงานที่อยู่ในแผนกที่ 9 จะถูกเปลี่ยนค่าเป็นค่าว่างก่อนหลังจากนั้นการลบข้อมูลของแผนกที่มีรหัส 9 จะถูกลบทิ้งหรือแก้ไขทันที ภายใน Entity แผนก

การลบหรือแก้ไขข้อมูลแบบใช้ค่าโดยปริยาย (Default) การลบหรือแก้ไขข้อมูลของคีย์หลัก สามารถทำได้โดยถ้าหากมีคีย์นอกที่อ้างอิงถึงคีย์หลักที่ถูกลบหรือแก้ไข ก็จะทำการปรับค่าของคีย์นอกนั้นโดยปริยาย (Default Value) ที่ถูกกำหนดขึ้นเช่น ในกรณีที่ยกเลิกแผนก 9 ใน Entity แผนกข้อมูลของพนักงานที่อยู่แผนก 9 ใน Entity พนักงานจะถูกเปลี่ยนค่าเป็น 00 ซึ่งเป็นค่าโดยปริยายที่หมายความว่าไม่ได้สังกัดแผนกใด ไม่เข้าใจกลับไปอ่านใหม่

Section 13: Data warehouse

13.1 วัตถุประสงค์การเรียนรู้

1) ทราบความหมายของ Data warehouse

13.2 Data warehouse

ข้อมูลกลายเป็นสินทรัพย์เชิงกลยุทธ์ที่ใช้ในการแปลงธุรกิจเพื่อค้นพบข้อมูลเชิงลึกใหม่ๆ ตามวิธีดังเดิมข้อมูลได้รับการรวบรวมในคลังข้อมูลขององค์กรซึ่งทำหน้าที่เป็นศูนย์กลางของข้อเท็จจริง อย่างไรก็ตามโลกของข้อมูลมีการพัฒนาอย่างรวดเร็วในรูปแบบที่เปลี่ยนอุตสาหกรรมและสร้างแรงจูงใจให้กับวิสาหกิจเพื่อพิจารณาแนวทางใหม่ในการเข้าถึงข้อมูลเชิงลึก นอกเหนือจากแหล่งข้อมูลดั้งเดิมจากระบบ ERP, CRM และ LOB แล้วแหล่งข้อมูลประเภทใหม่ๆ กำลังขับเคลื่อนการวิเคราะห์ที่แปรเปลี่ยนไปสู่ธุรกิจและมันมาจากข้อมูลที่สร้างขึ้นโดยทุกอย่างรอบตัวเรา เช่น แอปโซเชียลมีเดีย เว็บไซต์และอุปกรณ์ที่เชื่อมต่อ โดยรวมแล้ว IDC คาดการณ์ว่าการระเบิดข้อมูลนี้จะทำให้เกิดข้อมูลประมาณ 40 Zeta byte ในจักรวาลดิจิทัล ภายในปี 2563

ความท้าทายสำหรับองค์กรด้านไอทีคือคลังข้อมูลองค์กรแบบดั้งเดิมของตนไม่ได้ออกแบบมาเพื่อรวมการระเบิดของข้อมูลประเภทใหม่ลงในปริมาณมากๆ และรวดเร็ว เพื่อแก้ปัญหานี้จะต้องมีการเปลี่ยนแปลงอย่างมาก มีรายงานใน Gartner กล่าวว่า “คลังข้อมูลได้ถึงจุดที่ทำให้เกิดการเปลี่ยนแปลงที่สำคัญที่สุดตั้งแต่เริ่มก่อตั้ง สิ่งที่ใหญ่ที่สุดและอาจจะพิถีพิถันที่สุดในระบบจัดการข้อมูลในองค์กรไอที คือ การเปลี่ยนแปลง” (Roxane, Donald, & Merv, 2012) และเพื่อขับเคลื่อนธุรกิจไปข้างหน้า องค์กรสมัยใหม่จำเป็นต้องพัฒนาคลังข้อมูล (Data warehouse) ขององค์กรเพื่อให้สามารถใช้ประโยชน์จากข้อมูลขนาดใหญ่และดำเนินการได้แบบเรียลไทม์ เมื่อข้อมูลทั้งหมดถูกรวมไว้แล้วจะช่วยให้นักวิเคราะห์ธุรกิจและนักวิทยาศาสตร์ข้อมูลค้นพบข้อมูลเชิงลึกใหม่ๆ ที่ส่งผลกระทบต่อธุรกิจ การทำเช่นนี้คลังข้อมูลแบบดั้งเดิมต้องมีวิวัฒนาการไปสู่คลังข้อมูลที่ทันสมัย

13.3 คลังข้อมูลแบบดั้งเดิม

คลังข้อมูลแบบดั้งเดิมได้รับการออกแบบมาโดยเฉพาะเพื่อให้เป็นพื้นที่เก็บข้อมูลส่วนกลางสำหรับข้อมูลทั้งหมดใน บริษัท ข้อมูลที่แตกต่างจากระบบการทำธุรกรรม ERP CRM และ LOB แอปพลิเคชันจะถูกทำความสะอาด ได้แก่ การสกัด การเปลี่ยนรูป และการโหลด (ETL) เข้าไปในคลังข้อมูลและภายใน schema โดยรวม โครงสร้างข้อมูลคาดการณ์ได้และการประมวลผลที่ดีที่สุดและการรายงานผลการปฏิบัติงาน อย่างไรก็ตามการจัดเตรียมข้อความค้นหาส่วนใหญ่ได้รับการสนับสนุนทางด้าน IT และขึ้นอยู่กับการประมวลผลแบบ batch ตามตารางงานที่กำหนดไว้

Web 2.0 ได้เพิ่มข้อมูลทางธุรกิจที่เกี่ยวข้องกับข้อมูลที่สร้างขึ้นจากอีคอมเมิร์ซ เว็บบล็อก การค้นหาการตลาดและแหล่งข้อมูลอื่นๆ อย่างมาก แหล่งข้อมูลเหล่านี้ยังคงเป็นธุรกิจที่สร้างขึ้นและเป็นเจ้าของธุรกิจ องค์กรได้ขยายการดำเนินงาน ETL เพื่อชดเชยแหล่งข้อมูลใหม่ๆ และขยาย schema เพื่อรองรับสิ่งนี้

แม้ความซับซ้อนที่เพิ่มขึ้นเหล่านี้มูลค่าทางธุรกิจหลักของคลังข้อมูลแบบเดิมก็คือความสามารถในการวิเคราะห์และรายงานข้อมูลย้อนหลังจากแหล่งข้อมูลที่เชื่อถือได้และครบถ้วน ดังรูปที่ 13.1

รูปที่ 13.1 เฟรมเวิร์คของคลังข้อมูลแบบเดิม

(Rujirapong, ออกแบบ data lake platform อย่างไรให้สำเร็จ, 2017)

Section 14: NoSQL Data base

14.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจการนำฐานข้อมูล NoSQL ไปใช้ประโยชน์

14.2 NoSQL

NoSQL หรือ not only SQL เป็นเทคโนโลยีในการเรียกดูข้อมูลที่แตกต่างจาก SQL ซึ่งใช้ในฐานข้อมูลแบบสัมพันธ์ (relational database) โดยจุดเด่นของ NoSQL คือความง่ายในการขยายระบบที่เป็นรูปแบบ cluster และสามารถเรียกดูข้อมูลได้รวดเร็วจึงถูกนำมาใช้งานเกี่ยวกับข้อมูลขนาดใหญ่ และ real-time web application โดยหลักการของ NoSQL คือใช้โครงสร้างของข้อมูลได้หลายรูปแบบอาทิ key-value, wide column, graph หรือ document ซึ่งความสามารถของ NoSQL ถูกเพิ่มเข้ามาใน Web2.0 รูปแบบของ NoSQL มีหลากหลายรูปแบบและไม่มีมาตรฐานตายตัว ทั้งนี้ได้มีการจัดกลุ่มของ NoSQL ไว้คร่าว ๆ ดังเช่น Column, Document, Key-value, Graph, Multi-model ดังแสดงตัวอย่างฐานข้อมูลแบบ NoSQL รูปแบบต่างๆ ดังตารางที่ 14.1 โดยจะพบว่าฐานข้อมูลบางตัวถูกจัดไว้ในหลายกลุ่ม เนื่องจากมีความสามารถหลายด้าน

ตารางที่ 14.1 ตัวอย่างของฐานข้อมูล NoSQL

Column	Document	Key-Value	Graph	Multi-Model
Accumulo, Cassandra, Druid, HBase, Vertica, SAP HANA	Apache CouchDB, ArangoDB, Clusterpoint, Couchbase, DocumentDB, HyperDex, IBM Domino, MarkLogic, MongoDB, OrientDB, Qizx, RethinkDB	Aerospike, ArangoDB, Couchbase, Dynamo, FairCom c-treeACE, FoundationDB, HyperDex, InfinityDB, MemcacheDB, MUMPS, Oracle NoSQL Database, OrientDB, Redis, Riak, Berkeley DB	AllegroGraph, ArangoDB, InfiniteGraph, Apache Giraph, MarkLogic, Neo4J, OrientDB, Virtuoso, Stardog	Alchemy Database, ArangoDB, CortexDB, Couchbase, FoundationDB, InfinityDB, MarkLogic, OrientDB

จุดเด่นของฐานข้อมูลแบบ Key-Value คือมีความสามารถในการประมวลผลที่สูงมากเนื่องจากเป็นรูปแบบที่ง่ายที่สุด โดยทำงานในรูปแบบอาร์เรย์ (array) ซึ่งมีโครงสร้างของความสัมพันธ์เป็นคู่โดยที่ key จะต้องไม่ซ้ำกัน

สำหรับรูปแบบ Document มีจุดเด่นอยู่ที่การเข้ารหัสข้อมูลด้วยรูปแบบมาตรฐาน เช่น XML, YAML, JASON หรือ binary และเก็บในฐานข้อมูลด้วย key ชนิดไม่ซ้ำ (unique) ที่เชื่อมกับข้อมูลคล้ายกับรูปแบบของ key-value จุดเด่นอีกประการ คือ ในข้อมูลแต่ละชุด (record) สามารถมีจำนวนฟิลด์ไม่เท่ากันได้ซึ่งฐานข้อมูลแบบสัมพันธ์ซึ่งมีโครงสร้างเป็นตารางไม่สามารถทำได้ สำหรับภาษาที่ใช้ในการ query ข้อมูลสามารถติดตั้งได้หลายรูปแบบตามกลุ่มของ Document เช่น Collection, Tags, Non-visible metadata, Directory hierachies

ฐานข้อมูลชนิด graph เป็นฐานข้อมูลที่ถูกออกแบบมาสำหรับข้อมูลที่ต้องแสดงเป็นกราฟ หรือระบบโครงข่าย

Section 15: Big Data processing

15.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจกระบวนการวิเคราะห์ข้อมูลขนาดใหญ่

15.2 กระบวนการวิเคราะห์ข้อมูลขนาดใหญ่

ในการประมวลผลข้อมูลขนาดใหญ่จำเป็นต้องใช้กระบวนการทาง วิทยาการข้อมูล ซึ่งโดยทั่วไปกระบวนการดังกล่าวเริ่มต้นจากการตั้งคำถามง่ายๆ ที่สนใจ จนนำไปสู่การวิเคราะห์และแสดงผลลัพธ์เพื่อตอบคำถาม โดยขั้นตอนกระบวนการทาง วิทยาการข้อมูล จะขออ้างอิงจาก Blitzstein และ Hanspeter ซึ่งได้แสดงขั้นตอนไว้ดังรูปที่ 15.1 สามารถอธิบายได้ดังนี้

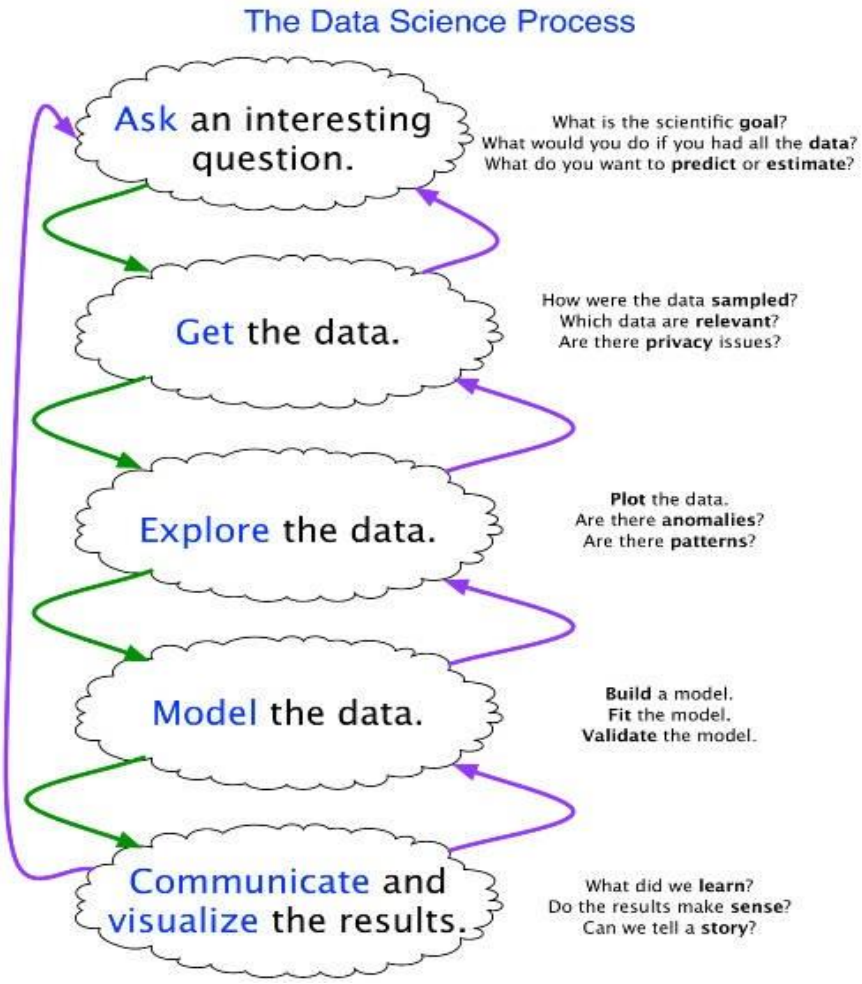
1. ตั้งคำถามที่น่าสนใจ (Ask an Interesting Question) เป็นกระบวนการแรกเริ่มของกระบวนทาง วิทยาการข้อมูล โดยคำถามนั้นไม่จำเป็นต้องซับซ้อนแต่ควรเป็นคำถามง่ายๆ ซึ่งจะเป็นตัวกำหนดทิศทางของกระบวนการถัดๆ ไป เช่น เป้าหมายคืออะไร จะทำอะไรจากข้อมูลที่มีอยู่ หรือ ต้องการที่จะคาดการณ์หรือประเมินผลในเรื่องใดบ้าง เป็นต้น

2. เก็บข้อมูล (Get the Data) เป็นขั้นตอนของการเก็บรวบรวมข้อมูลทั้งแบบที่มีโครงสร้างและไม่มีโครงสร้าง หลังจากที่เราทราบปัญหาและรู้ว่าหากต้องการตอบปัญหานั้นจะต้องใช้ข้อมูลอะไรบ้างและมาจากแหล่งข้อมูลใดบ้าง

3. สำรวจข้อมูล (Explore the Data) เป็นการตรวจสอบข้อมูลรวมถึงการแปลงข้อมูลที่ได้จากขั้นตอนการเก็บข้อมูลให้เหมาะสมก่อนที่จะนำเข้าสู่ขั้นตอนการวิเคราะห์ข้อมูล โดยในขั้นตอนนี้อาจใช้การ plot ข้อมูลเพื่อดูรูปแบบของข้อมูล ข้อมูลมีความผิดปกติหรือไม่ ทั้งนี้ข้อมูลที่ผ่านในกระบวนการนี้แล้วต้องมีความเหมาะสมสอดคล้องกับแบบจำลองที่เราต้องการจะใช้ในการวิเคราะห์ข้อมูล

4. สร้างแบบจำลองเพื่อการวิเคราะห์ข้อมูล ในขั้นตอนนี้ถือไปขั้นตอนที่สำคัญและนักวิทยาการข้อมูล จะต้องมีทักษะในการเลือกใช้แบบจำลองที่หลากหลายและเหมาะสมกับข้อมูลอินพุต รวมไปถึงต้องสอดคล้องกับโจทย์ปัญหาที่ต้องการหาคำตอบ ซึ่งหากนักวิทยาการข้อมูล ไม่มีทักษะที่ดีก็จะทำให้เลือกใช้แบบจำลองข้อมูลที่ไม่เหมาะ สมส่งผลให้ของการวิเคราะห์ข้อมูลไม่ถูกต้องตามไปด้วย ซึ่งโดยทั่วไปแล้วแบบจำลองที่ว่านี้ส่วนใหญ่จะใช้วิธีการทาง machine learning เข้ามาแก้ปัญหา เช่น Decision tree, Association rule, Artificial neural networks, Support vector machines, Bayesian networks, Genetic algorithms เป็นต้น ซึ่งไม่สามารถบอกได้ว่าวิธีการใดดีที่สุดในการหาคำตอบขึ้นอยู่กับข้อมูลและโจทย์ปัญหา

5. การสื่อสารและการแสดงผลลัพธ์ หลังจากที่ผ่านขั้นตอนของการวิเคราะห์แล้ว ขั้นตอนสุดท้ายของกระบวนการวิทยาการข้อมูล คือการนำผลจากการวิเคราะห์มาแสดงผลให้อยู่ในรูปแบบที่เข้าใจได้ง่ายสามารถสื่อสารให้คนที่เกี่ยวข้องในแต่ละระดับให้เข้าใจและสามารถนำไปใช้งานได้



รูปที่ 15.1 กระบวนการทางวิทยาการข้อมูลถูกนำเสนอโดย Blitzstein และ Hanspeter

(Mark & Douglas, 2012)

Section 16: Big Data Architecture and Analytics Platforms with Hadoop's architecture

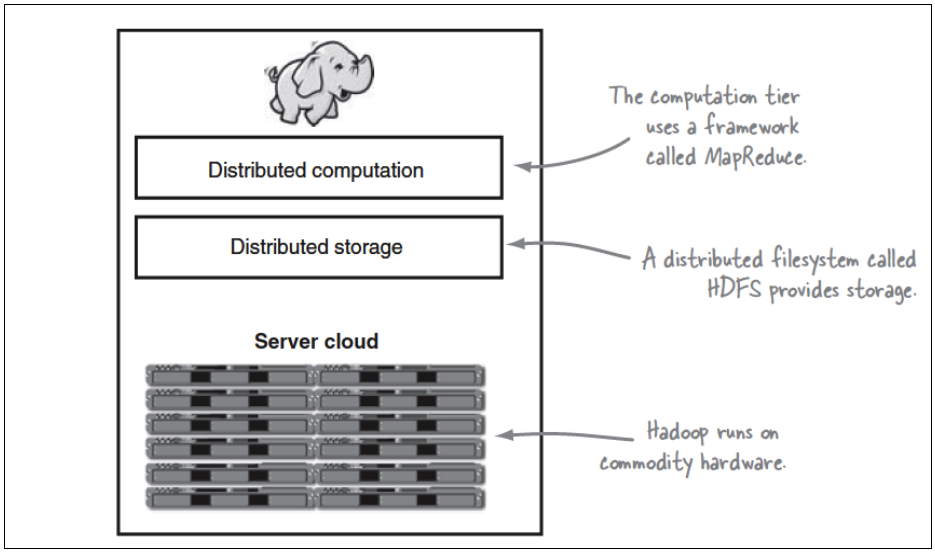
16.1 วัตถุประสงค์การเรียนรู้

1) เข้าใจสถาปัตยกรรมของ Hadoop
2) เข้าใจระบบนิเวศของ Hadoop

16.2 Hadoop

ตามนิยามคุณลักษณะของข้อมูลขนาดใหญ่ ด้วย 3V: Volume, Variety และ Velocity นั้นเครื่องมือในการทำ การวิเคราะห์ข้อมูลขนาดใหญ่ ก็จะต้องเปลี่ยนไปจากที่เคยใช้ RDBMS ที่เป็น SQL ต้องเปลี่ยนเป็นเครื่องมืออื่นๆ ที่สามารถจัดการข้อมูลได้จำนวนมากขึ้นอย่าง NewSQL เช่น MySQL Cluster, Amazon RDS หรือ Azure SQL หรือเครื่องมือที่เป็น NoSQL อย่าง MongoDB หรือ Cassandra และเครื่องมืออย่าง Hadoop ที่ได้รับความสนใจอย่างกว้างเพราะสามารถที่จะจัดการข้อมูล Unstructured ขนาดใหญ่ได้ เช่นข้อมูลที่ เป็น Text File, XML หรือ JSON

Hadoop เป็น Open source Project ของ Apache สำหรับการเก็บและบริหารข้อมูลขนาดใหญ่ (Holmes, 2012) Hadoop เขียนด้วยโปรแกรมภาษาจาวา มีความสามารถในการทำ Fault Tourarent เพราะจะเก็บข้อมูลซ้ำกันใน หลายๆ ที่ และเป็นระบบที่เป็น Horizontal Scale ที่รันบนเครื่อง commodity server จำนวนมาก Hadoop Project เริ่มต้นโดย Doug Cutting และ Mike Cafarella ที่เป็นทีมงานของบริษัท Yahoo ซึ่งต่อมาก็มีบริษัทอื่นๆ นำไปใช้กันอย่างกว้างขวางเช่น eBay, Facebook และ Amazon รวมถึงมีบริษัทหลายๆ รายที่นำ Hadoop มาทำ Commercial Distribution อาทิเช่น Cloudera, MapR, IBM Infoshphere BigInsight, Hortonwork ดังรูปที่ 16.1 สภาพแวดล้อมของ Hadoop

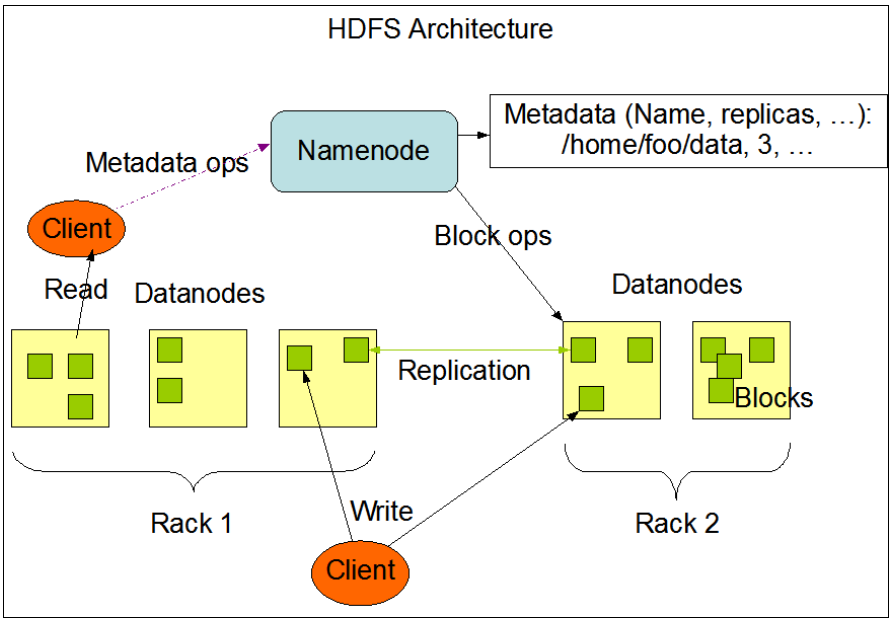


รูปที่ 16.1 สภาพแวดล้อมของ Hadoop

(Forrester, 2016)

Hadoop จะมีองค์ประกอบหลักสองส่วนคือ

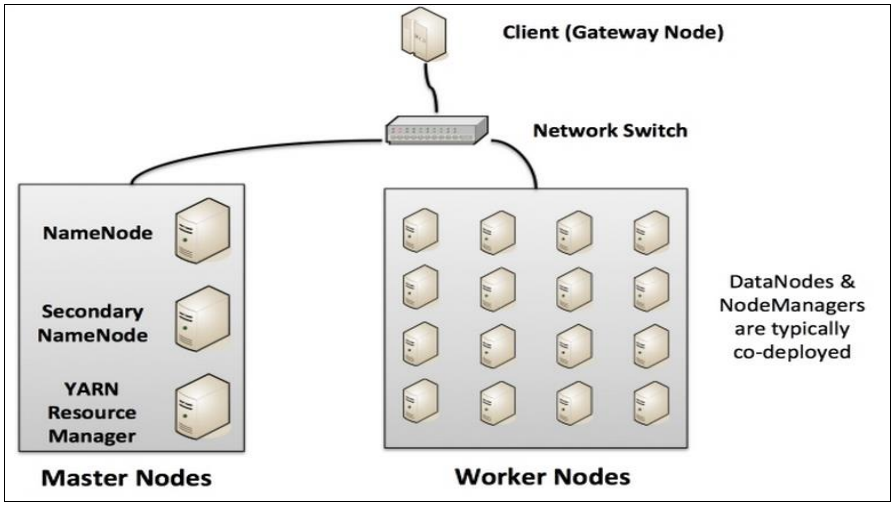
- HDFS (Hadoop Distribution File System) ทำหน้าที่เป็นส่วนเก็บข้อมูล ซึ่งจะเก็บข้อมูลขนาดใหญ่ที่ถูกแบ่งเป็นไฟล์ย่อยเก็บลง ใน Data Node จำนวนมาก โดยจะมี Master Node ทำหน้าที่ระบุตำแหน่ง ของข้อมูลที่เก็บใน Data node ดังรูปที่ 16.2



รูปที่ 16.2 สถาปัตยกรรมของ Hadoop HDFS

(Forrester, 2016)

ตามรูปที่ 16.3 การทำ Hadoop Cluster ถือว่าเป็นสิ่งจำเป็นในการประมวลผลข้อมูลขนาดใหญ่ เนื่องจากต้องใช้ทรัพยากรร่วมกันอย่างมาก



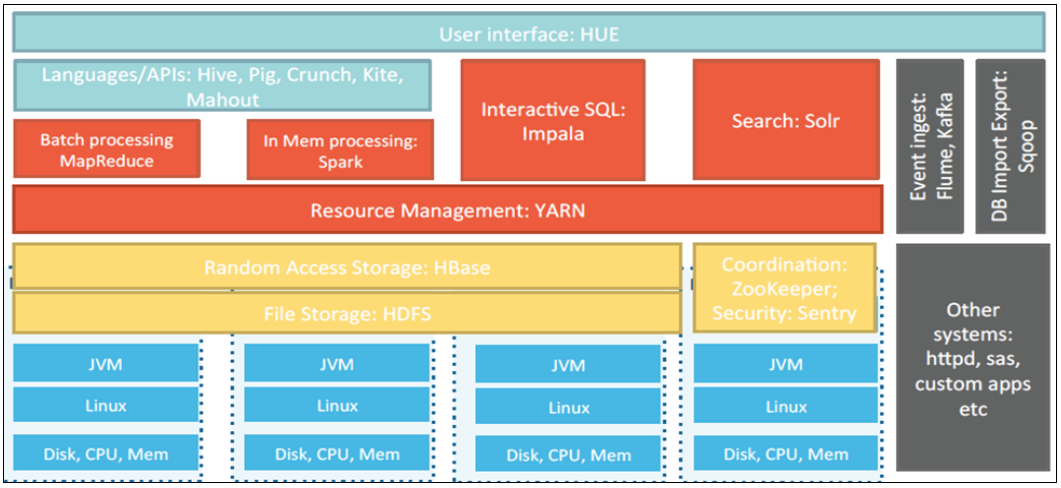
รูปที่ 16.3 การทำ Hadoop Cluster

(Forrester, 2016)

16.3 ระบบนิเวศของ Hadoop

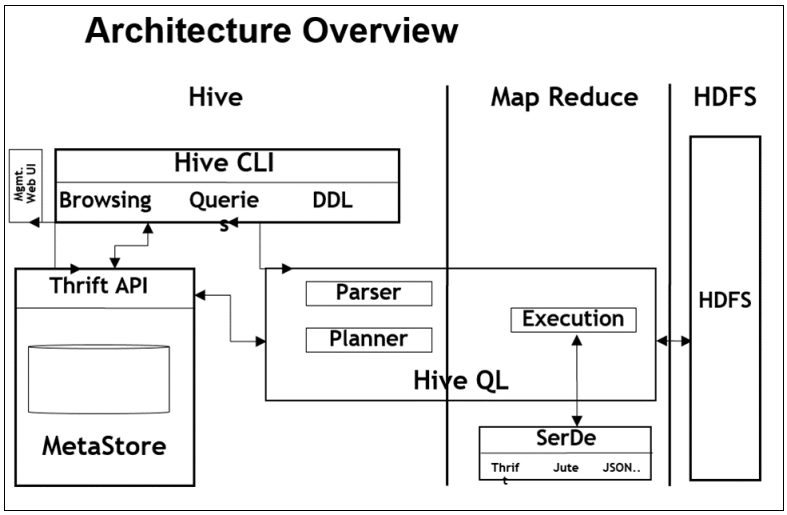
ระบบ Hadoop เองจะมีองค์ประกอบหลักอยู่แค่สองส่วนคือ HDFS และ Map/Reduce ซึ่งค่อนข้างจะไม่สะดวกกับผู้ใช้งานที่มีความต้องการอื่น เช่น การประมวลผลโดยใช้ภาษา SQL การเขียนหรืออ่านข้อมูลแบบ Random access หรือการถ่ายโอนข้อมูลจากที่อื่น จึงมีการพัฒนาโครงการที่มาทำงานร่วมกับ Hadoop โดยแนวคิดระบบนิเวศของ Hadoop Zoo แสดงได้ดังรูปที่ 16.4 กรอบแนวคิดระบบนิเวศของ Hadoop เพื่อให้ได้ประสิทธิภาพดียิ่งขึ้น ซึ่งมีเครื่องมือที่สำคัญดังนี้

- Hive เป็นเครื่องมือสำหรับผู้ต้องการสืบค้น (Query) ข้อมูลที่เก็บใน HDFS ด้วยภาษาลักษณะ SQL แทนที่จะต้องมาเขียนโปรแกรม Map/ Reduce โดย Hive จะทำหน้าที่ในการแปล SQL line ให้มาเป็น Map/Reduce แล้วก็ทำการรันแบบ Batch ดังรูปที่ 16.5 สถาปัตยกรรมของ Hive



รูปที่ 16.4 กรอบแนวคิดระบบนิเวศของ Hadoop

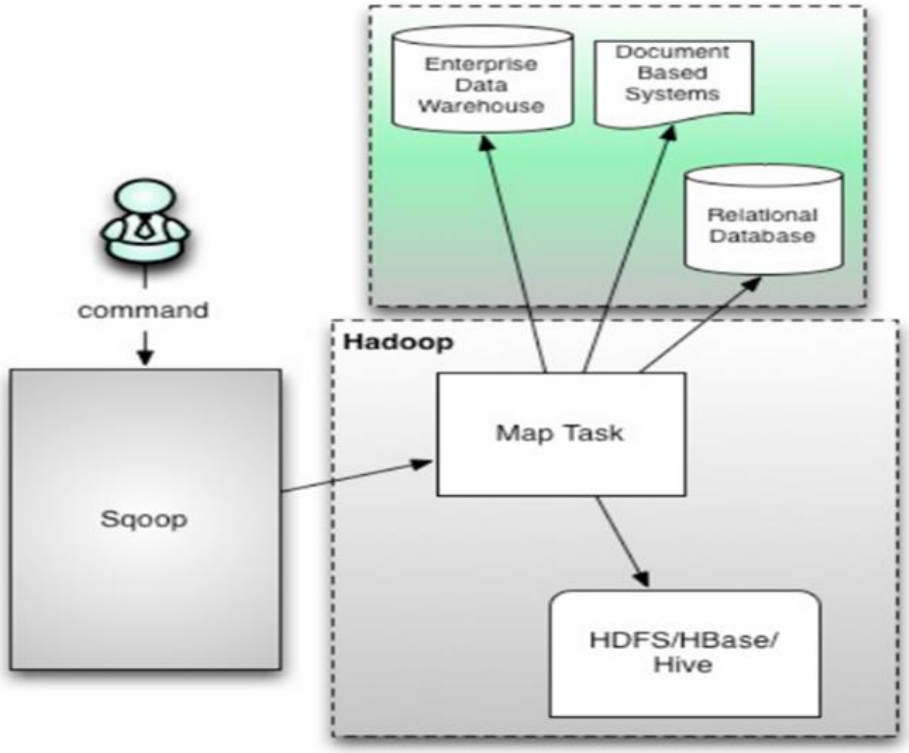
(Forrester, 2016)



รูปที่ 16.5 สถาปัตยกรรมของ Hive

(Forrester, 2016)

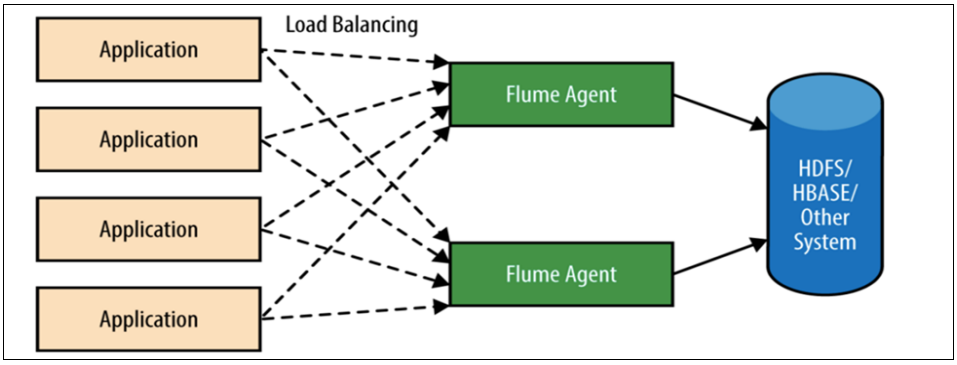
- Pig เป็นเครื่องมือคล้ายๆกับ Hive ที่ช่วยให้ประมวลผลข้อมูลโดยไม่ต้องเขียนโปรแกรม Map/Reduce ซึ่ง Pig จะใช้โปรแกรมภาษา script ง่าย ๆ เรียกว่า Pig Latin แทน โดย Pig เหมาะกับการทำ ETL สำหรับการแปลงข้อมูลในรูปแบบต่าง ๆ เช่น JSON
- Sqoop เป็นเครื่องมือในการถ่ายโอนข้อมูลระหว่างฐานข้อมูลที่อยู่รูปแบบ Table บน RDBMS อย่าง SQL server, Oracle หรือ MySQL กับข้อมูลบน HDFS ของ Hadoop ดังรูปที่ 16.6 สถาปัตยกรรมของ Sqoop



รูปที่ 16.6 สถาปัตยกรรมของ Sqoop

(Forrester, 2016)

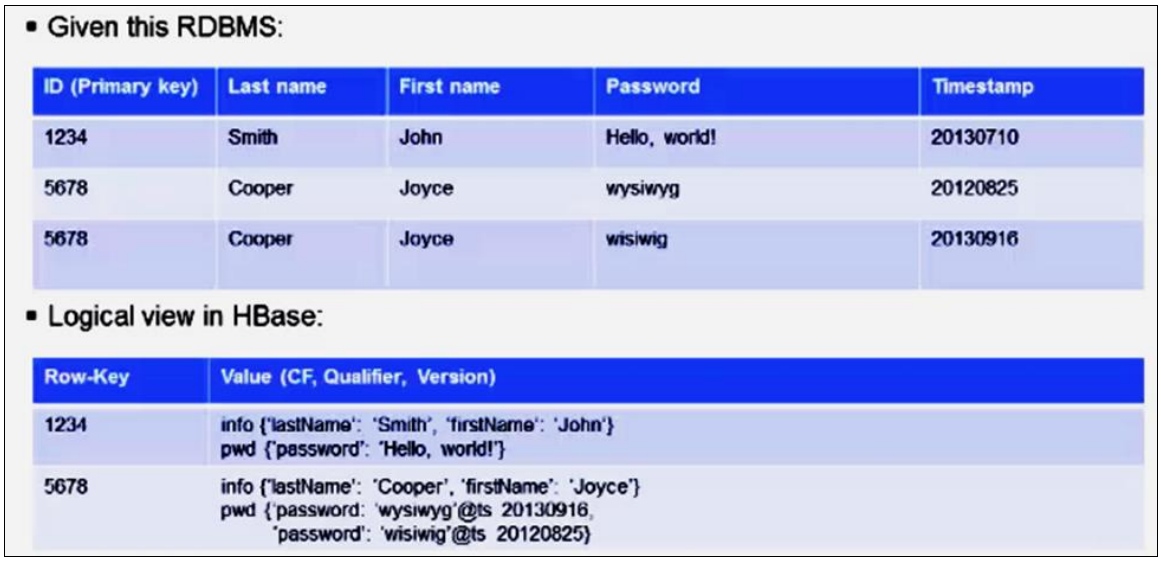
- Flume เป็นเครื่องมือในการดึงข้อมูลจากระบบอื่น ๆ แบบ real-time เข้าสู่ HDFS เช่นการดึง Log จาก Web Server การดึงข้อมูลเหล่านี้จะต้อง มีการติดตั้ง Agent ที่เครื่อง Server ดังรูปที่ 16.7



รูปที่ 16.7 การทำงานของ Flume

(Forrester, 2016)

- HBase เป็นเครื่องมือที่จะทำให้ Hadoop สามารถอ่านและเขียนข้อมูล แบบ real-time Random Access ได้โดยจะทำให้เป็น BigTable ที่เก็บ ข้อมูลได้ไม่จำกัด row หรือ column ซึ่ง HBase ก็จะเป็นเสมือนการทำให้ Hadoop เป็น NoSQL Database ดังรูปที่ 16.8 ลักษณะข้อมูลใน RDBMS และ HBase



- Given this RDBMS:

ID (Primary key)	Last name	First name	Password	Timestamp
1234	Smith	John	Hello, world!	20130710
5678	Cooper	Joyce	wysiwyg	20120825
5678	Cooper	Joyce	wisiwig	20130916

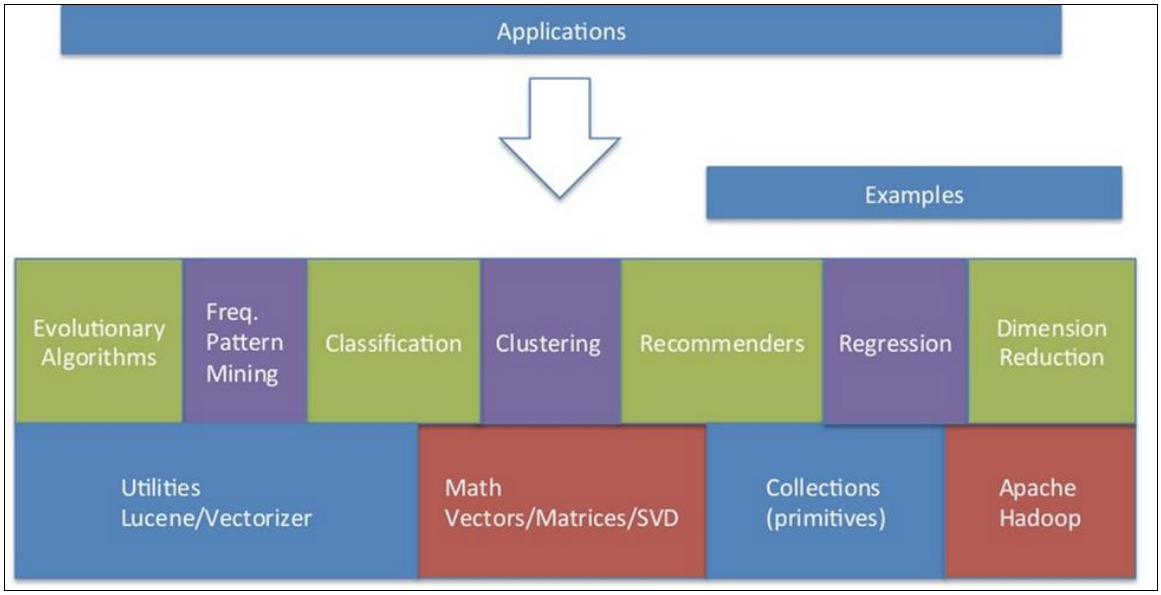
- Logical view in HBase:

Row-Key	Value (CF, Qualifier, Version)
1234	info {'lastName': 'Smith', 'firstName': 'John'} pwd {'password': 'Hello, world!'}
5678	info {'lastName': 'Cooper', 'firstName': 'Joyce'} pwd {'password: 'wysiwyg'@ts 20130916, 'password': 'wisiwig'@ts 20120825}

รูปที่ 16.8 ลักษณะข้อมูลใน RDBMS และ HBase

(Holmes, 2012)

- Oozie เป็นเครื่องมือในการทำ Workflow จะช่วยให้เราเอาคำสั่ง ประมวลผลต่างๆ ของระบบ Hadoop เช่น Map/Reduce, Hive หรือ Pig มาเชื่อมต่อกันในรูปของ Workflow ได้
- Hue ย่อมาจากคำว่า Hadoop User Experience เป็นเครื่องมือช่วย ทำ User interface ของ Hadoop ให้ใช้งานได้ง่ายขึ้นกว่าการต้องใช้ command line



รูปที่ 16.9 สถาปัตยกรรมของ Mahout

(Holmes, 2012)

- Mahout เป็นเครื่องมือของ Data Scientist ที่ต้องการทำ Predictive Analytics ข้อมูลบน Hadoop โดยใช้ภาษาจาวา ทั้งนี้ Mahout สามารถใช้ Algorithm ที่เป็น Recommender, Classification และ Clustering ได้ ดังรูปที่ 16.9 สถาปัตยกรรมของ Mahout

เอกสารอ้างอิง

Powell, G., & et al. (2016). Social media listening for routine post-marketing safety surveillance. *Drug safety*(39.5), 443-454.

Azizan, S. A., & Aziz, I. A. (2017). Terrorism Detection Based on Sentiment Analysis Using Machine Learning. *Journal of Engineering and Applied Sciences, 12*(3), 691-698.

Concordiam, P. (2014). Extremism hits home stopping the spread of terrorism. *J. Eur. Secur. Defense*(5), 1-67.

ECMA International. (2013). *The JSON Data Interchange Format.*

Cearley, D. W., Burke, B., & Walker, M. J. (2016). *Top 10 Strategic Technology Trends for.*

Holmes, A. (2012). *Hadoop in practice.* Manning Publications Co.

Data Science Central. (2018). *Data Science Central*. Retrieved April 8, 2018, from https://www.datasciencecentral.com/

Petersen, R. (2016). *37 Big Data Case Studies with Big Results*. Retrieved March 28, 2018, from Schaefer Marketing Solutions: https://www.businessesgrow.com/2016/12/06/big-data-case-studies/

IGCSE ICT. (2018). *Expert Systems IGCSE ICT*. Retrieved April 14, 2018, from IGCSE ICT: https://www.igcseict.info/theory/7_2/expert/

IDCs Big Data and Analytics Maturity Assessment. (2016). Retrieved March 25, 2018, from csc.bigdatamaturity: http://csc.bigdatamaturity.com/

Kvangundy. (2015, October 13). Retrieved April 11, 2018, from Kvangundy: http://kvangundy.com/wp/sentiment-analysis-amazon-reviews-using-neo4j/

Microsoft. (2012, April). *Microsoft Case Studies, Bank of Nagoya Dramatically Accelerates Database Queries and Increases Availability*. Retrieved April 9, 2018, from Microsoft Case Studies: http://www.microsoft.com/casestudies/Case_Study_Detail.aspx?CaseStudyID=710000000344

Microsoft. (2010, November). *Microsoft Case Studies, Stock Exchange Chooses Windows over Linux; Reduces Latency by 83 Percent*. Retrieved April 1, 2018, from http://www.microsoft.com/casestudies/Windows-Server-2008-R2-Enterprise/Direct-Edge/Stock-Exchange- Chooses-Windows-over-Linux-Reduces-Latency-by-83-Percent/4

RightScale. (2015). *See the Latest Cloud Computing Trends*. Retrieved April 9, 2018, from 2015 State of the Cloud Report: http://www.rightscale.com/lp/2015-state-of-the-cloud-report?campaign=701700000012UP6

Microsoft. (2013, May). *Stock Exchange Gains Deeper Understanding of Data and Drives New Business Growth*. Retrieved April 12, 2018, from Microsoft Case Studies: http://www.microsoft.com/casestudies/Microsoft-Excel-2010/Direct-Edge/Stock-Exchange-Gains-Deeper- Understanding-of-Data-and-Drives-New-Business-Growth/710000002540

Mathworks. (2018). *What is Deep Learning?* Retrieved April 3, 2018, from Mathworks: https://www.mathworks.com/discovery/deep-learning.html

hooktalk. (2017, January 12). *การทำ Social Listening เมืองไทย*. Retrieved March 24, 2018, from http://hooktalk.com

InfoMobius. (2015). *InfoMobius*. Retrieved April 7, 2018, from BIG DATA ช่วยเพิ่มคุณค่าให้ธุรกิจได้อย่างไร : http://www.infomobius.com/2015/03/how-big-data-delivers-business-values/

Adam, G., & Josh, P. (2017). *Deep Learning.* O'Reilly Media, Inc. Retrieved April 14, 2018, from https://www.safaribooksonline.com/library/view/deep-learning/9781491924570/ch04.html

Amazon. (2018). *what is a data lake*. Retrieved April 5, 2018, from https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Arun, M. (2017, July 5). *Kernel SVM (Support Vector Machine)*. Retrieved April 14, 2018, from http://arun-aiml.blogspot.com/2017/07/kernel-svm-support-vector-machine.html

Braun, H. T. (2015). *Evaluation of Big Data Maturity Models: A benchmarking study to support big data assessment in organizations.* Tampere University of Technology, Faculty of Business and Built Environment. Retrieved from https://dspace.cc.tut.fi/dpub/bitstream/handle/123456789/23016/braun.pdf?sequence=1

Cao, D.-Z., Pang, S.-L., & Bai, Y.-H. (2005). Forecasting exchange rate using support vector machines. *International Conference on Machine Learning and Cybernetics. 6*, pp. 3448-3452. Guangzhou: IEEE.

Chatdanai, L. (2017, November 20). *รู้จักกับ self-organizing map*. Retrieved April 10, 2018, from TUA on Github: https://clumdee.github.io/blog/self-organizing-map/

Evans, R. G. (2015, June 2). Retrieved March 20, 2018, from greenlightdigital: http://www.greenlightdigital.com/blog/posts/why-you-should-have-a-social-media-listening-strategy/

Forrester. (2016). *Big Data Hadoop Cloud Solutions, Q2 2016*. Retrieved April 2, 2018, from The Forrester WaveTM: The Forrester WaveTM: Big Data Hadoop Cloud Solutions, Q2 2016 https://www.forrester.com/report/The+Forrester+Wave+Big+Data+Hadoop+Cloud+Solutions+Q2+2016/-/E- RES126541#figure4

Fowler, M. (2015, February 5). *Data Lake*. Retrieved April 1, 2018, from https://martinfowler.com/bliki/DataLake.html

Fukunaga, K. (1990). *Statistical Pattern Recognition 2nd Edtion.* Academic Press.

Gartner. (2013, September 12). *Big Data Business Benefits Are Hampered by Culture Clash*. Retrieved April 3, 2018, from Gartner: https://www.gartner.com/doc/2588415/big-data-business-benefits-hampered

Halper, F., & Stodder, D. (2016). *TDWI Benchmark Guide: A Guide to Achieving Big Data Analytics Maturity.* TDWI.

Halper, F., & Krishnan, K. (2013). *TDWI Big Data Maturity Model Guide: Interpreting your assessment score.* TDWI Research. Retrieved from https://tdwi.org/whitepapers/2013/10/tdwi-big-data-maturity-model-guide/asset.aspx?tc=assetpg&tc=page0&tc=assetpg

InfoTech.com. (2013, September 25). *Big Data Maturity Assessment Tool*. (Infotech) Retrieved April 11, 2018, from https://www.infotech.com/research/ss/leverage-big-data-by-starting-small/it-big-data-maturity-assessment-tool#more-details

Jarupreechachan, R. (2016). *7 โมเดลธุรกิจที่จะทำให้ Big Data กลายเป็น Big profit*. Retrieved March 20, 2018, from http://bigdataexperience.org

Mark, A. B., & Douglas, L. (2012, June 21). *The Importance of "Big Data": A Definition*. Retrieved April 10, 2018, from Gartner: http://www.gartner.com/id=2057415

Microsoft. (2012, May). *Hy-Vee Boosts Performance, Speeds Data Delivery, and Increases Competitiveness*. Retrieved April 1, 2018, from Microsoft Case Studies: http://www.microsoft.com/casestudies/Microsoft-SQL-Server-2008-R2-Enterprise/Hy-Vee/Hy-Vee-Boosts-Performance-Speeds-Data-Delivery-and-Increases-Competitiveness/710000000776

Nott, C. (2014, August 15). *IBM: Big Data & Analytics Hub*. Retrieved March 18, 2018, from http://www.ibmbigdatahub.com/blog/big-data-analytics-maturity-model

Phattaraphon, O. (2017, November 22). Retrieved April 13, 2018, from medium.com: https://medium.com/artificialintelligence06/sentiment-analysis-กับเครื่องดักฟังชั้นเยี่ยม-570cb4d1b66a

Radcliffe, J. (2014, January 7). *Radliffe Advisory Services.* Retrieved February 27, 2018, from www.radcliffeadvisory.com/research/download.php?file=RAS_BD_MatMod.pdf

Roxane, E., Donald, F. A., & Merv, A. (2012, February 13). *The State of Data Warehousing in 2012*. Retrieved April 1, 2018, from https://www.gartner.com/id=1922714

Rujirapong, R. (2017, October 17). *Data Lake แนวคิดใหม่ที่ทุกองค์กรต้องเริ่มต้น*. Retrieved April 1, 2018, from https://rujirapong.wixsite.com/softnix/single-post/2017/10/17/Data-Lake

Rujirapong, R. (2018, February 20). *สรุปความเข้าใจเกี่ยวกับ data lake*. Retrieved April 1, 2018, from https://medium.com/softnix/สรุปความเข้าใจเกี่ยวกับ-data-lake-คืออะไร-ต่างจาก-data-warehouse-ยังไง

Rujirapong, R. (2017, November 3). *ออกแบบ data lake platform อย่างไรให้สำเร็จ*. Retrieved April 2018, from https://medium.com/softnix/การออกแบบ-data-lake-platform-อย่างไรให้สำเร็จ

Sharma, G. (2016, November 19). Retrieved March 23, 2018, from paralleldots: https://blog.paralleldots.com/product/understanding-sentiment-analysis-use-cases/

Smith, O. (2016). *Social Listening/Monitoring*. Retrieved March 21, 2018, from Social Media Marketing: https://www.contentshifu.com/social-media-marketing/introduction-social-listening-monitoring/

Songyot, N., & David P., C. (2007). Adaptive branch and bound algorithm for selecting optimal features. *Science Direct, Pattern Recognition Letters, 28*, 1415-1427.

Veenstra, A. F. (2013). Big data in small steps: Assessing the value of data. *TNO*.

Wannaphong. (2017, February). *ทำ Sentiment Analysis ภาษาไทยใน Python*. Retrieved April 1, 2018, from https://python3.wannaphong.com/2017/02/sentiment-analysis-python.html

Willy, W. (2014). *Artificial Neural Networks and Pattern Recognition For students of HI.* Retrieved from Slideshare.

Google. (2019). *How Google Search Works*. Retrieved from
https://support.google.com/webmasters/answer/70897?hl=en