ณรงค์ฤทธิ์ มโนมัยพิบูลย์

“Big Data” is (now) all around

คำว่า “Big Data” เกิดขึ้นมาพร้อมๆ กับการเฟื่องฟูของสื่อสังคมออนไลน์ (Social Media) ซึ่งนำขบวนมาโดยเฟสบุ๊ค (Facebook) อย่างไรก็ตามคำว่า “Big Data” นั้นมีอะไรมากกว่าการมีปริมาณข้อมูลจำนวนมากตามชื่อเรียก คำว่า “Big Data” สามารถเชื่อมโยงไปถึงระบบการประมวลผลข้อมูลประเภทนี้ซึ่งค่อนข้างใหม่ และแตกต่างจากเทคโนโลยีเดิมที่มีใช้กันอย่างแพร่หลายในบ้านเรา

ซึ่งผมจะได้กล่าวต่อไปในบทความนี้ “Big Data” ยังเปิดประตูไปสู่ระบบงานทางด้านธุรกิจที่เกี่ยวข้องกับข้อมูลประเภทนี้โดยตรงไม่ว่าจะเป็นระบบงานที่ตอบโจทย์เฉพาะทางธุรกิจนั้นๆ หรือจะเป็นระบบงานบริการทางด้านข้อมูล ระบบวิเคราะห์ข้อมูล

คุณลักษณะที่สำคัญ 3 ประการของ Big Data คือ Volume Variety และ Velocity

- **Volume:** ปริมาณข้อมูลจะมีขนาดใหญ่ตั้งแต่ระดับ Terabytes Petabytes ไปจนถึง Zettabytes
- **Variety:** ชนิดของข้อมูลที่มีความหลากหลายไม่ว่าจะเป็น Structured, Semi-Structured หรือจะเป็น Unstructured Data โดยเฉพาะ Unstructured Data ซึ่งเป็นชนิดข้อมูลที่ถูกพูดถึงพร้อมกับ “Big Data”
- **Velocity:** มีการให้ความสำคัญ ความน่าสนใจกับข้อมูลประเภท “Real-Time” อย่างมากว่าจะนำมาใช้ให้เกิดประโยชน์ได้อย่างไร แต่ก็ไม่ได้ละเลยข้อมูลประเภทอื่น

ความจำเป็นของ “Big Data” ต่อธุรกิจ

“Big Data” อาจเป็นเรื่องที่ถูกกล่าวถึงมากในช่วงนี้ แต่ไม่ใช่ทุกองค์กรที่ต้องกังวลเกี่ยวกับการเกิดหรือการมาของยุค “Big Data” ตราบใดที่องค์กรสามารถดำเนินธุรกิจต่อไปได้ เว้นเสียแต่ว่า Big Data จะเป็นแหล่งข้อมูลของธุรกิจโดยตรงที่จะขาดไม่ได้ ยกตัวอย่าง สื่อสังคมออนไลน์ถูกจัดว่าเป็นแหล่งข้อมูลแบบ “Big Data” ประเภทหนึ่ง ซึ่งสื่อประเภทนี้มีความจำเป็นต่อธุรกิจที่ต้องพึ่งกิจกรรมทางการตลาดสูง โดยเป็นทั้งผู้ให้และผู้รับข้อมูลจากสื่อ แต่สื่อประเภทนี้มีความสำคัญน้อยมากกับโรงงานอุตสาหกรรม

ในทางตรงข้าม ข้อมูลอีกกลุ่มหนึ่งที่ถูกจัดว่าเป็นแหล่งข้อมูลทางด้าน “Big Data” ที่สำคัญ และแต่ละธุรกิจให้ความสำคัญเป็นอันดับต้นๆ คือ ข้อมูลจาก “ระบบบันทึกการเปลี่ยนแปลง” (Logs หรือ Transaction Logs) ของระบบงานต่างๆ รวมถึงข้อมูลจากระบบเซ็นเซอร์ (Censors) ต่างๆ ซึ่งมีลักษณะสำคัญคือเกิดขึ้นได้ตลอดเวลาที่ระบบทำงาน และสามารถบ่งชี้สถานะในแง่ปริมาณและคุณภาพของการทำงานในระบบได้ อาทิ ระบบ ATM (Automatic Teller Machine) ในธุรกิจธนาคารและสถาบันการเงิน หรือระบบ CDR (Call Detail Records) สำหรับธุรกิจโทรคมนาคม ระบบควบคุมการผลิต (Shop Floor Control) สำหรับธุรกิจทางด้านอิเล็กทรอนิกส์ ระบบ Smart Metering สำหรับธุรกิจโครงสร้างพื้นฐาน (ไฟฟ้า น้ำประปา) เป็นต้น ข้อมูลกลุ่มนี้มีผลต่อสินค้า และบริการของผู้ประกอบการธุรกิจโดยตรง ความท้าทายในส่วนนี้ก็คือ ผู้ประกอบการจะสามารถหาประโยชน์หรือคุณค่าที่ซ่อนอยู่ ด้วยวิธีการประมวลผลข้อมูลกลุ่มนี้ได้อย่างมีประสิทธิผลและประสิทธิภาพเพียงใด เพื่อรับรู้สถานการณ์ป้องกันปัญหา แก้ปัญหาให้ทันท่วงที นอกจากจะหวังผลในคุณค่าของข้อมูลที่ซ่อนอยู่แล้ว การประมวลผลต้องมีความรวดเร็วตั้งแต่ระดับวันต่อวัน ชั่วโมงต่อชั่วโมง หรือวินาทีต่อวินาทีเลยทีเดียว ซึ่งถึงตอนนี้คงไม่มีใครปฏิเสธว่า การประมวลผลข้อมูลในรูปแบบนี้เป็นจริงได้ และบางระบบกำลังให้บริการพวกเราอยู่โดยที่เราไม่รู้ตัว

เทคโนโลยีสำหรับประมวลผล “Big Data”

ดังที่ได้เกริ่นนำในตอนต้นว่า “Big Data” ยังเชื่อมโยงไปสู่ระบบการประมวลผลสำหรับข้อมูลปริมาณมาก ผมขอจัดแบ่งเทคโนโลยีที่จะมีบทบาทใน “Big Data” ออกเป็น 4 กลุ่มดังนี้

- **เทคโนโลยีหลัก**ที่ถือว่าอยู่เบื้องหลัง “Big Data” คือ “Hadoop” ซึ่งเป็นซอฟต์แวร์แบบโอเพ่นซอร์ส (Open-source Software) ของ Apache สำหรับการประมวลผลแบบกระจาย หรือ Distributed Computing เพื่อรองรับการจัดเก็บ และประมวลข้อมูลขนาดใหญ่ Hadoop ได้รวมระบบการจัดการเครื่องแม่ข่ายในลักษณะคลัสเตอร์ และการเข้าถึงและดึงข้อมูลอย่างรวดเร็วด้วยวิธี MapReduce (Map และ Reduce) จากความสามารถข้างต้นของ Hadoop ระบบคอมพิวเตอร์ที่จะรองรับการทำงานของ Hadoop จะเป็นกลุ่มเครื่องแม่ข่ายขนาดเล็กหลายๆ เครื่องมีหน่วยจัดเก็บข้อมูลภายในขนาดใหญ่ในแต่ละเครื่อง (ปัจจุบันมีหน่วยจัดเก็บข้อมูลภายนอกมาเป็นทางเลือกแล้ว) ต่อเชื่อมกันผ่านระบบเครือข่าย (Local Area Network) หรือเครือข่ายระยะไกล (Wide Area Network)

นอกจากนี้ยังมีพันธมิตรของ Hadoop หรือที่เรียกว่า Hadoop Ecosystem อีกจำนวนหนึ่งที่จะมาช่วยเสริมในเรื่องการจัดการข้อมูล การเข้าถึงและดึงข้อมูล รวมทั้งการติดต่อแลกเปลี่ยนข้อมูลกับระบบต่างๆ ให้สะดวกขึ้น อาทิ HBase, Hive, Pig, Sqoop เป็นต้น เห็นชื่อแล้วคงไม่ค่อยคุ้นกัน เพราะทั้งหมดนี้เป็นซอฟต์แวร์แบบโอเพ่นซอร์สทั้งหมด โดยมี Hadoop เป็นแกนกลางในการทำงาน

องค์กรสามารถดาวน์โหลด Hadoop และผลิตภัณฑ์อื่นในกลุ่ม Hadoop Ecosystem มาใช้งานได้โดยไม่มีค่าใช้จ่าย และเพื่อตอบโจทย์การนำ Hadoop มาใช้ในธุรกิจ จึงมีบริษัทซอฟต์แวร์ที่ตั้งขึ้นมาเพื่อทำหน้าที่ให้บริการทางด้าน Hadoop Ecosystem แบบครบวงจรตั้งแต่อำนวยความสะดวกในการดาวน์โหลด ไปจนถึงการสนับสนุนหลังการดาวน์โหลด ปัจจุบันมีบริษัทที่ทำหน้านี้อยู่ 4 แห่งคือ Cloudera (CDH), MapR, Hortonworks และบริษัทน้องใหม่อย่าง Pivotal HD

- **เทคโนโลยีกลุ่มที่สอง**คือ ระบบฐานข้อมูลที่ไม่ใช้ภาษา SQL (NoSQL Database) เนื่องจากความสามารถที่รวดเร็ว สามารถรองรับข้อมูลแบบ Semi-Structured และ Unstructured ได้ ผลิตภัณฑ์ที่นิยมใช้ส่วนใหญ่เป็นโอเพ่นซอร์ส และรองรับการขยายตัวในแนวราบ (Horizontal Scaling) ซึ่งสอดคล้องกับสถาปัตยกรรมของ Hadoop ตัวอย่างผลิตภัณฑ์ทางด้าน NoSQL Database ที่เป็นที่นิยมได้แก่ Cassandra, CouchBase, HBase, MongoDB เป็นต้น
- **เทคโนโลยีกลุ่มที่สาม**คือ “Data Visualization Tools” ซึ่งเป็นเครื่องมือที่จะช่วยแปลงข้อมูล “Big Data” ที่ได้รับการกลั่นกรองแล้วมาแสดงในรูปของแผนภาพ ง่ายต่อการเข้าใจ และนำไปสู่การตัดสินใจในขั้นถัดไป แล้วเครื่องมือกลุ่มนี้ต่างจากระบบ Business Intelligence อย่างไร บทบาทของเครื่องมือกลุ่มนี้จะอยู่ในระดับปฏิบัติการ (Operations) ให้ติดตามสถานะของระบบ และการแก้ปัญหาได้ง่าย โดยมีคำเรียกสำหรับระบบนี้ว่า “Operational Intelligence” ส่วน Business Intelligence จะเน้นไปที่ข้อมูลสำหรับผู้บริหาร ผู้จัดการเพื่อประกอบการตัดสินใจทางธุรกิจ
- **เทคโนโลยีกลุ่มสุดท้าย**คือ “Analytic Database” ผลิตภัณฑ์ในกลุ่มนี้อาจจะนำไปใช้กับระบบคลังข้อมูลได้ด้วย

และเป็นกลุ่มผู้ผลิตซอฟต์แวร์ยักษ์ใหญ่ในตลาดต่างให้ความสำคัญมาก โดยใช้เทคนิคในการทำงานแบบต่างๆ เพื่อตอบโจทย์ด้านความเร็วไม่ว่าจะเป็น การประมวลผลในหน่วยความจำ (In-memory Computing) การประมวลในระบบฐานข้อมูล (In-database Computing) ซึ่งไม่เหมือนกันเลยแต่มีสิ่งหนึ่งที่ทุกผู้ผลิตมีเหมือนกันคือ การสนับสนุนการต่อเชื่อมกับ Hadoop เพื่อให้สามารถนำข้อมูลจาก Hadoop เข้ามาประมวลในขั้นต่อไปในผลิตภัณฑ์ฐานข้อมูลของตนเองได้ ซึ่งเกือบทุกผู้ผลิตจะมีการนำ Hadoop เข้ามาเป็นผลิตภัณฑ์เสริมของตนเองโดยทำสัญญากับทางบริษัทที่ให้การสนับสนุน Hadoop Ecosystem ทั้ง 4 รายข้างต้น ตัวอย่างผลิตภัณฑ์ในกลุ่มนี้ได้แก่ Aster Data (Teradata), Exadata (Oracle), Greenplum (EMC) Netezza (IBM), Vertica (HP) เป็นต้น

ตัวอย่างข้อมูลอ้างอิงเกี่ยวกับ “Big Data” และระบบที่ใช้ในการบริหารจัดการข้อมูล

- จากข้อมูลในเดือนกรกฎาคม ปี 2552 Facebook มีเครื่องแม่ข่ายที่มีหน่วยประมวลผลรวมกว่า 4800 CPU Cores และหน่วยจัดเก็บข้อมูลหลักอีก 2 Petabytes เพื่อสนับสนุนระบบคลังข้อมูลที่ใช้ Hadoop และ Hive เป็นหลัก ขณะที่ในปีเดียวกัน Yahoo ใช้หน่วยประมวลผลรวมกว่า 32,000 CPU Cores หน่วยความจำหลัก (RAM) รวมมากกว่า 64 Terabytes และหน่วยจัดเก็บข้อมูลมากกว่า 16 Petabytes ในระบบ Hadoop ซึ่งเติบโตเป็นเท่าตัวจากปี 2551

“Big Data” กับการเปลี่ยนแปลงในระบบงานทางด้านธุรกิจ

ในส่วนของระบบงาน (Applications) “Big Data” มีส่วนทำให้เกิดการเปลี่ยนแปลงอย่างมากเช่นกัน ซึ่งผมขอแบ่งออกเป็น 3 ลักษณะ ได้แก่

- การเปลี่ยนแปลงในขั้นตอนเกี่ยวกับการนำข้อมูลเข้าระบบงานที่กำลังใช้อยู่ (Data Integration) โดยการเพิ่มช่องทางในการนำข้อมูลจากแหล่งใหม่โดยเฉพาะ “Big Data” ซึ่งไม่ได้กระทบต่อตัวระบบงาน อาทิ ในขั้นตอนจัดเตรียมข้อมูลเพื่อนำเข้าระบบคลังข้อมูล มีการเพิ่มส่วนในการดึงข้อมูลจาก Hadoop หรือ NoSQL Database เป็นต้น
- การพัฒนาระบบงานเดิมด้วยวิธีการหรือเทคโนโลยีใหม่ เริ่มตั้งแต่การนำการประมวลผลและการจัดการข้อมูลแบบกระจายของ Hadoop มาใช้เพื่อให้การทำงานเร็วขึ้น ในต่างประเทศมีการพัฒนาระบบคลังข้อมูล (DataWarehouse) โดยใช้กลุ่มผลิตภัณฑ์ใน Hadoop Ecosystem ทั้งหมดเพื่อรองรับลักษณะของข้อมูลที่หลากหลายของธุรกิจ นอกจากนี้มีการพูดถึงระบบ “Big Data Analytics” ซึ่งเน้นไปที่การวิเคราะห์หาคุณค่าจากข้อมูลประเภทนี้ด้วยวิธีการใหม่
- การพัฒนาระบบงานเพื่อธุรกิจแต่ละประเภท (Industry-based) จะสามารถทำได้ลึกและรวดเร็วขึ้นโดยอาศัยข้อมูลดิบเดิมที่มีอยู่ ผมขอยกกรณีศึกษามาเพื่อให้เห็นภาพระบบงานที่เกี่ยวข้องกับ “Big Data” กรณีศึกษาแรกเป็นกรณีศึกษาที่แพร่หลายมาก คือ ระบบ “Customer Churn Analysis” สำหรับธุรกิจโทรคมนาคม คือการวิเคราะห์พฤติกรรมของผู้บริโภคในการใช้บริการโทรศัพท์เคลื่อนที่โดยการนำข้อมูลจากระบบ CRM (Customer Relationship Management) มาเพื่อใช้ในการวิเคราะห์หาลูกค้าที่มีแนวโน้มว่าจะเปลี่ยน หรือลูกค้าที่ต้องรักษาเอาไว้ การจัดเตรียมข้อเสนอพิเศษต่างๆ กรณีศึกษาที่สองคือ ทีมรถแข่ง F1 ของแม็คลาเรน (McLaren) ได้นำข้อมูลจากเซ็นเซอร์ของรถแข่งในระหว่างการแข่งขันมาเข้าระบบ Predictive Analysis ในการประเมินปัญหาที่เกิดขึ้นกับรถแข่ง เพื่อให้สามารถแก้ไขปัญหาได้ทันท่วงที

ระบบงานอีกประเภทหนึ่งที่จะเติบโตในยุคของ Big Data คือ Electronic Discovery (e-Discovery) หรือจะเรียกว่า Information Discovery เนื่องจากความต้องการระบบที่สามารถสืบค้นข้อมูลต่างๆ ที่มีความสัมพันธ์กัน (Correlated Information) จากข้อมูลปริมาณมาก จากหลายแหล่งข้อมูลในเวลาที่รวดเร็ว ความพิเศษของระบบงานประเภทนี้คือสามารถสืบค้นข้อมูลได้หลายประเภท (Structured, Semi-Structured หรือ Unstructured) และสามารถสืบค้นข้อมูลตามสื่อสังคมออนไลน์ชั้นนำต่างๆ ได้ ในอเมริกามีธุรกิจที่นิยมใช้ระบบนี้มากคือ ที่ปรึกษาทางกฎหมาย โดยใช้รวบรวม จัดเตรียมข้อมูลเพื่อเป็นหลักฐานในการฟ้องร้องคดีในศาล ส่วนองค์กรที่อาศัยการตลาดเป็นตัวนำจะใช้ระบบนี้เพื่อติดตามสถานะของสินค้า บริการ หรือแม้แต่ภาพพจน์ของตนเอง จากความเห็นของผู้บริโภคทั้งในแง่บวกและลบ

ความท้าทายขององค์กรในประเทศไทย

หลายองค์กรในต่างประเทศให้ความเห็นตรงกันว่าเรื่องบุคลากรมีความสำคัญที่สุด จากการคอยเป็นผู้บริโภคข่าวสารข้อมูล จะเปลี่ยนเป็นผู้ใช้ข้อมูลให้เกิดประโยชน์ บทบาทหน้าที่ที่เคยแบ่งไปตามของความรู้ความชำนาญแต่ละคน จะเปลี่ยนมารวมในคนๆ เดียวมากขึ้น นั่นคือพนักงานจำเป็นต้องมีความรู้ความชำนาญหลายแขนงมากขึ้น โดยเฉพาะความสามารถในการวิเคราะห์ข้อมูล บุคลากรต้องได้รับการอบรม เตรียมตัวเพื่อให้สามารถสนับสนุนเทคโนโลยีใหม่อย่างเช่น Hadoop, NoSQL มีตำแหน่งงานใหม่ๆ ได้เกิดขึ้นในต่างประเทศแล้ว ได้แก่ Data Scientists, Big Data Engineers มหาวิทยาลัยชั้นนำในต่างประเทศได้จัดเตรียมหลักสูตรสำหรับผลิตบุคลากรทางด้าน Data Science มากขึ้น (Oracle ให้ข้อมูลว่า Data Scientist จะต้องมีพร้อมทั้งความรู้ทางด้านธุรกิจ ควบคู่ไปกับความชำนาญด้านการวิเคราะห์ นั่นคือ Industry Expert บวกกับ Analytic Skills)

สุดท้ายนี้ “Big Data” ไม่ได้ให้ผลลัพธ์อะไรที่ใหม่หรือซับซ้อนไปกว่า ข้อมูลที่ผ่านการวิเคราะห์ (Data Analytics) และแสดงผลในรูปแบบต่างๆ (Data Visualization) ที่ง่ายต่อการเข้าใจ โดยการนำข้อมูลปริมาณมากมาผ่านการประมวลผลการวิเคราะห์ และแสดงผลด้วยวิธีที่เหมาะสม