รายงานสรุปการอบรม/สัมมนา/พัฒนาความรู้/ประชุมเชิงปฏิบัติการ/และเป็นวิทยากร
กองนโยบายและแผนการใช้ที่ดิน กรมพัฒนาที่ดิน

****

ส่วนที่ 1 ข้อมูลทั่วไป

ชื่อ นายดิเรก นามสกุล คงแพ
ตำแหน่ง นักวิเคราะห์นโยบายและแผนชำนาญการพิเศษ กลุ่ม/ฝ่าย กลุ่มวางแผนบริหารจัดการพื้นที่ชุ่มน้ำ
หลักสูตร/หัวข้อเรื่องอบรม/สัมมนา/พัฒนาความรู้ฯ
หลักสูตร Uses of Hadoop in Big Data : เครื่องมือในการวิเคราะห์ข้อมูล
สถานที่อบรม/สัมมนา/พัฒนาความรู้ฯ
ระบบการฝึกอบรมผ่านสื่ออิเล็กทรอนิกส์ (https://e-learning.dga.or.th/xlms_ega/resource/tincan/c97a3fda-ec19-412d-8647-daec81ed29dd/index.html)
หน่วยงานที่จัดฝึกอบรม/สัมมนา/พัฒนาความรู้ฯ
สถาบันพัฒนาบุคลากรภาครัฐด้านดิจิทัล (Thailand Digital Government Academy)
ตั้งแต่วันที่ 30 เดือน มกราคม พ.ศ. 2565 ถึงวันที่ 31 เดือน มกราคม พ.ศ. 2565
เพื่อ ☑ อบรม ☐ สัมมนา ☐ อื่นๆ ระบุ.....

ส่วนที่ 2 สิ่งที่ได้รับจากการอบรม/สัมมนา/พัฒนาความรู้

2.1 รายงานสรุปเนื้อหาสาระสำคัญในการอบรม/ สัมมนา/พัฒนาความรู้ฯ

การอบรม Uses of Hadoop in Big Data : เครื่องมือในการวิเคราะห์ข้อมูล แนะนำ Big Data เรียนรู้เกี่ยวกับหลักการทำงานของฮาดูปและระบบที่เกี่ยวข้องกับฮาดูปเพื่อการวิเคราะห์ข้อมูลใหญ่เบื้องต้น ความรู้เกี่ยวกับ Big Data ความหมายและลักษณะที่สำคัญของ Big Data รวมทั้งแนวทางการวิเคราะห์ด้วยเครื่องมือต่าง ๆ และแนวทางการนำ Big Data ไปใช้ประโยชน์เพื่อการทำงาน มีวัตถุประสงค์ 1) เพื่อให้ผู้เรียนมีความรู้เกี่ยวกับเครื่องมือต่าง ๆ ที่ใช้ในการวิเคราะห์ข้อมูล 2) เพื่อให้ผู้เรียนสามารถเลือกใช้เครื่องมือในการวิเคราะห์ข้อมูลเพื่อการทำงานได้อย่างเหมาะสม

อธิบายแยกตามหลักสูตร แบ่งได้เป็น

1) ทำความรู้จัก Big data เป็นคำศัพท์ใหม่ที่เพิ่งเริ่มใช้ในช่วงปี 1990 โดยมี John Mashey ผู้ที่ทำให้คำนี้เป็นที่รู้จักขึ้นมา ซึ่งปกติแล้ว Big Data จะเป็นข้อมูลที่มีปริมาณที่ใหญ่มากโดยที่ซอฟท์แวร์รุ่นเก่าไม่สามารถประมวลผลได้ หรือสามารถประมวลผลได้แต่ใช้เวลานาน โดย Big Data จะมีทั้งข้อมูลที่มีโครงสร้าง กึ่งมีโครงสร้าง และไม่มีโครงสร้าง ขนาดของ Big Data ก็จะมีการเพิ่มขนาดของข้อมูลขึ้นไปเรื่อยๆ จากปัจจุบันที่มีขนาดข้อมูลหลายพัน Terabytes ก็จะมีการเพิ่มขนาดเป็น Zettabytes
การทำงานของ Big Data ต้องอาศัยเทคนิคและเทคโนโลยีสมัยใหม่ที่สามารถรองรับข้อมูลที่มีขนาดใหญ่ ซับซ้อน และหลากหลายได้ โดยในปี 2018 ได้มีการนิยาม Big Data ใหม่ว่า "Big Data คือ เครื่องมือที่ใช้ในการจัดการกับข้อมูล"

ลักษณะสำคัญของ Big Data

Big Data จะต้องมีลักษณะสำคัญ 4V ดังต่อไปนี้ จึงจะเรียกได้ว่าเป็น Big Data

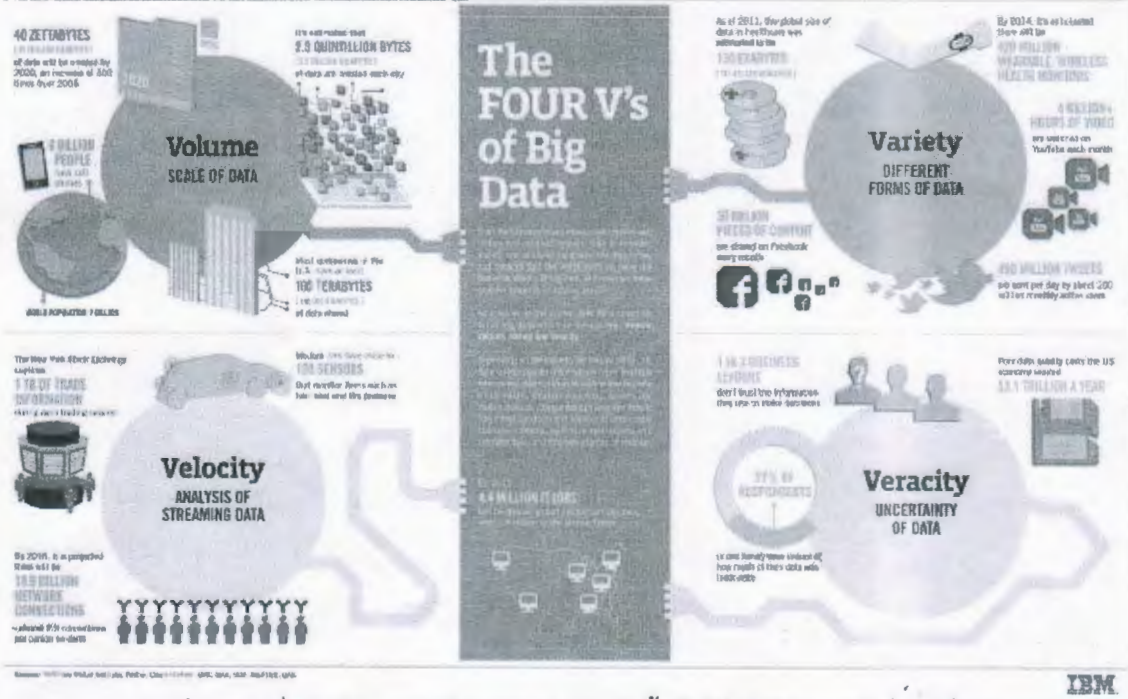
(1) ปริมาณ (Volume) คือ ปริมาณข้อมูลที่สามารถผลิตและจัดเก็บไว้ได้จะต้องขนาดที่ใหญ่มากเพียงพอ ซึ่งปริมาณของข้อมูลจะเป็นข้อบ่งบอกได้ถึงคุณภาพและประสิทธิภาพของข้อมูลภายในนั้น และสามารถนำไปพิจารณาต่อได้ว่าข้อมูลที่มีอยู่เหล่านั้น ถือเป็น Big Data หรือไม่

(2) ความหลากหลาย (Variety) คือ ความหลากหลายของประเภทของข้อมูล โดยสามารถเพิ่มประสิทธิภาพในการนำข้อมูลไปวิเคราะห์ต่อยอดได้ ไม่ว่าจะเป็นข้อมูลที่เป็นตัวหนังสือ, รูปภาพ, ข้อมูลเสียงที่ถูกบันทึกไว้, วีดีโอหรือไฟล์ประเภทอื่นจากหลากหลายแหล่งที่มา ก็สามารถเป็นส่วนหนึ่งในการเพิ่มประสิทธิภาพการทำงานของ Big Data ได้ ซึ่งจะเป็นข้อมูลจากทุกฝ่ายไม่ว่าจะเป็นการเงิน ฝ่ายขาย การตลาด ลูกค้าสัมพันธ์ บัญชี รวมถึงตัวหนังสือที่อาจเป็นบทสนทนาระหว่างแผนก หรือระหว่างบริษัท ซึ่งอาจเป็นข้อความจาก Social Media รวมถึง URLs ที่มีข้อมูลเข้ามาอยู่ในทุกวัน

(3) ความเร็ว (Velocity) คือ ความเร็วในการประมวลผลและผลิตข้อมูลขึ้นมาเพื่อให้ทันกับความต้องการของผู้ใช้งาน ซึ่ง Big Data คือข้อมูลที่ได้มาแบบ Real-Time และประมวลผลอยู่ตลอดเวลา แตกต่างจาก Small Data ที่ไม่สามารถทำได้ Big Data จะมีความถี่ในการประมวลผลที่มากกว่า การบันทึกข้อมูลที่มากกว่า และเผยแพร่ข้อมูลที่มากกว่า ข้อมูลของ Big Data ที่มีอัตราการเพิ่มขึ้นของข้อมูลที่มีการเพิ่มขึ้นได้อย่างรวดเร็ว โดยไม่ว่าจะเป็น

-ข้อมูลตัวอักษรการสนทนา
-ข้อมูลการบันทึกเสียง
-ข้อมูลการถ่ายภาพวีดีโอ
-ข้อมูลอัตราการสั่งซื้อสินค้า
-ข้อมูลโปรโมชั่นต่างๆ
-ซึ่งหากมานั่งดูแล้วจะถือว่าเป็นข้อมูลที่มีอยู่มาก และมีการอัปเดทเคลื่อนไหวอยู่ตลอดเวลา

(4) คุณภาพของข้อมูล (Veracity) คือ คุณภาพของข้อมูลที่สามารถนำไปวิเคราะห์ข้อมูลต่อไปได้อีก เป็นข้อมูลที่ยังไม่ผ่านการประมวลผลอยู่ในรูปแบบของข้อมูลดิบซึ่งสามารถนำไปประมวลผลต่อไปได้ และเป็นข้อมูลที่มาจากหลากหลายแหล่งไม่ว่าจะเป็น Facebook, Youtube, Twitter ซึ่งข้อมูลจากแหล่งเหล่านี้ยากที่จะสามารถควบคุมคุณภาพรวมถึงการคัดกรองข้อมูล และความน่าเชื่อถือของข้อมูล จึงต้องนำข้อมูลเหล่านี้เข้าสู่กระบวนการทำ Data Cleansing



นอกจาก 4V ที่เป็นลักษณะสำคัญของ Big Data นี้แล้วก็ยังมีลักษณะข้ออื่นๆ ที่สามารถบ่งบอกได้ว่าเป็น Big Data เช่นเดียวกัน ได้แก่

Scalability คือ ขนาดของข้อมูลทั้งหมดที่มีที่ต้องสามารถขยายขนาดได้อย่างรวดเร็ว

Relational คือ ความเกี่ยวข้องกันของข้อมูล ข้อมูลที่มีความเกี่ยวข้องกันอยู่จะสามารถทำให้การประมวลผลสามารถทำได้ดีมากยิ่งขึ้น

Big Data เป็นเครื่องมือที่ช่วยให้สามารถใช้ข้อมูลได้อย่างมีประสิทธิภาพ ข้อมูลจากทุกแหล่งที่มาสามารถนำไปวิเคราะห์และวางแผนเพื่อให้ได้ผลลัพธ์ตามที่ต้องการส่วนหนึ่ง เพื่อช่วยให้สามารถเข้าถึงความต้องการของผู้บริโภคได้มากที่สุด เมื่อลดระยะเวลาในการดำเนินงานได้ ก็สามารถลดต้นทุนได้เป็นอย่างดี ซึ่งประโยชน์ในส่วนนี้ของ Big Data ทำให้บริษัทใหญ่ๆ หลายเจ้านำไปใช้ในการวางแผนการตลาดเพื่อวิเคราะห์ลักษณะของผู้บริโภครวมถึงความต้องการของผู้บริโภค สำหรับรองรับธุรกิจขนาดใหญ่เช่น บริษัทขายของออนไลน์ ใช้ Big Data ในการวิเคราะห์พฤติกรรมของลูกค้าโดยอาศัยข้อมูลจากการ Tracking (ติดตามพฤติกรรมการใช้งาน) การค้นหาข้อมูลของลูกค้า ว่ามีความต้องการเป็นอย่างไร โดยยังสามารถตรวจสอบราคาของคู่แข่ง จำนวนสินค้า เพื่อนำเสนอให้ตรงกับความต้องการของลูกค้ามากที่สุด แล้วจึงนำข้อมูลเหล่านี้มาเสนอให้กับลูกค้าแบบอัตโนมัติและปรับราคาขึ้นหรือลงโดยอิงจากความต้องการของตลาด

กระบวนการทำงานของ Big Data

ขั้นตอนและกระบวนการทำงานของ Big Data มีอยู่ 3 ขั้นตอนหลักๆ ดังนี้

(1) จัดเก็บข้อมูล (Storage) เป็นขั้นตอนการจัดเก็บรวบรวมข้อมูลทั้งหมดจากแหล่งต่างๆ ไม่ว่าจะเป็นข้อมูลที่มีคุณภาพรวมถึงข้อมูลที่คาดว่าอาจจะเป็นประโยชน์ ไม่ว่าจะเป็นข้อมูลที่เป็นตัวอักษร ไฟล์เอกสาร ไฟล์รูปภาพ ไฟล์วีดีโอ ไฟล์เสียงที่ถูกบันทึก จะถูกเก็บรวบรวมไว้ที่นี่

(2) การประมวลผลข้อมูล (Processing) การประมวลผลข้อมูล หลังจากที่นำข้อมูลมารวบรวมไว้ได้ในที่เดียวแล้ว ข้อมูลต่างๆ จะถูกนำไปจัดหมวดหมู่ให้อยู่ในกลุ่มที่มีความเกี่ยวข้องกันหรือความสัมพันธ์ใกล้เคียงกัน ให้ผลลัพธ์คล้ายคลึงกันมากที่สุด แล้วจึงนำมาเปลี่ยนเป็นรูปแบบข้อมูลเพื่อนำเอาข้อมูลที่มีอยู่เหล่านี้เข้าระบบข้อมูลที่ผ่านการประมวลผลแล้ว

(3) การวิเคราะห์ข้อมูล (Analyst) การวิเคราะห์ข้อมูลและนำเสนอข้อมูล หลังจากที่ข้อมูลทั้งหมดได้ถูกจัดกลุ่มและแยกประเภทเรียบร้อยแล้วนั้น ต่อจากนั้นจะนำมาวิเคราะห์หา Pattern ความเกี่ยวข้องกันทั้งหมด ที่อาจมองไม่เห็นได้เลยด้วยตา ไม่ว่าจะเป็นการหา แนวโน้มของการตลาด ความต้องการของลูกค้า กระแสที่อาจเกิดขึ้นได้ในอนาคต และข้อมูลด้านอื่นที่เป็นประโยชน์ และจัดมานำเสนอในรูปแบบต่างๆ ไม่ว่าจะเป็นรูปภาพหรือกราฟ

การประยุกต์ข้อมูลขนาดใหญ่ จริงๆ แล้ว Big Data สามารถนำไปประยุกต์ใช้ได้กับหลายภาคส่วน ไม่ว่าจะเป็นภาครัฐ หรือภาคเอกชน ในส่วนนี้จะพูดถึงประโยชน์ของ Big Data หากนำมาปรับใช้ในภาคธุรกิจแล้วจะสามารถทำให้สามารถเข้าใจลูกค้าได้มากยิ่งขึ้น ด้วยการนำฐานข้อมูลที่มีอยู่ใน Big Data ศึกษาถึงลักษณะพฤติกรรมของผู้บริโภคว่ามีการตัดสินใจในการเลือกสินค้าอย่างไร สามารถนำเสนอสินค้าที่คุณมีอยู่ให้ตรงกับความต้องการของลูกค้าได้หรือไม่ หากไม่มีจะสามารถนำสินค้าชนิดอื่นที่มีอยู่นำไปทดแทนได้หรือไม่ สามารถวิเคราะห์ความต้องการของตลาดในอนาคตได้ ข้อมูลต่างๆ ที่ถูกค้นหาในอินเตอร์เน็ต รวมถึงใน Social Media ต่างๆ สามารถนำมารวบรวมได้ ว่ามีอะไรที่เป็นกระแสหรือได้รับความนิยมอยู่ในขณะนั้น ว่ากระแสอะไรที่นำมาค้นหาหรือกล่าวถึงอยู่มากที่สุด ก็สามารถนำข้อมูลนั้นมาวิเคราะห์และวางแผนก่อน หากมีแผนที่ดีและสามารถทำได้อย่างรวดเร็วก็สามารถเป็นผู้นำกระแสได้อย่างไม่ยาก จากการใช้ข้อมูลจาก Big Data การวางแผนในอนาคตมีประสิทธิภาพมากยิ่งขึ้น จากข้อมูลที่มีอยู่สามารถนำผลวิเคราะห์จาก Big Data เข้ามาช่วยประกอบการวางแผนและการตัดสินใจได้เป็นอย่างดี ทั้งในเรื่องของการลดต้นทุน เพิ่มผลผลิต โดยการเก็บข้อมูลต่างๆ ภายในองค์กรเพื่อนำไปวิเคราะห์ไม่ว่าจะเป็น กระบวนการผลิต ข้อมูลการใช้วัตถุดิบ จะทำให้สามารถทราบได้ว่าปัญหาภายในองค์กรมีหรือไม่ต้องปรับปรุงแก้ไขส่วนใด เพื่อที่จะสามารถแก้ไขปัญหาได้อย่างรวดเร็ว เพื่อป้องกันความผิดพลาดในการผลิต คาดการณ์ปัญหาที่อาจเกิดขึ้น จากการนำข้อมูลที่มีจาก Big Data มาคาดการณ์ความต้องการของตลาด ซึ่งนอกจากคาดการณ์ในอนาคตได้แล้วนั้น ก็ยังสามารถนำข้อมูลส่วนนั้นมาวิเคราะห์ต่อยอดได้อีกว่า ในอนาคตนั้นจะมีเหตุการณ์อะไรที่สามารถเกิดขึ้นได้บ้าง ก็

สามารถนำข้อมูลส่วนนั้นนำไปวางแผน ปรับนโยบาย วิธีการบริหารองค์กร เพื่อให้สามารถแก้ไขปัญหาได้อย่างรวดเร็วที่อาจเกิดขึ้นได้ในอนาคต ลดงบประมาณในการลงทุนด้าน IT ในอนาคตหลังจากที่นำ Big Data มาใช้ในองค์กรแล้วสามารถลดต้นทุนการใช้งบประมาณในด้าน IT ได้เป็นอย่างดี เนื่องจากสามารถนำข้อมูลที่ได้มาไปใช้ประโยชน์ในด้านอื่นๆ ได้อีกพร้อมยังช่วยลดต้นทุนในการจ้างพนักงานในด้าน IT ได้อีกหลายตำแหน่งที่สามารถใช้ Big Data มาทดแทนได้

2) หลักการทำงานของ Hadoop

ฮาดูป (Hadoop)

ประวัติความเป็นมาของ Hadoop ต้องย้อนกลับไปในปี 2006 หลังจากที่ World Wide Web เติบโตจนถึงจุดที่การใช้งานอินเตอร์เน็ตมีการขยายวงกว้างออกไปเรื่อยๆ ผู้คนค้นหาข้อมูลต่างๆพอๆกับที่มีการป้อนคอนเท้นท์และข้อมูลเข้าไป ในปีนั้นเองที่ Google เริ่มมีการทำงานเกี่ยวกับการจัดเก็บข้อมูลและการประมวลผลข้อมูล Yahoo และทีมผู้พัฒนาซอฟท์แวร์จึงได้มีการเริ่มต้นพัฒนา Hadoop ขึ้น ซึ่งชื่อนี้มีที่มาจากชื่อของเล่นของลูกชายหัวหน้าทีมผู้พัฒนานั้นเอง จากนั้นในปี 2008 Yahoo ก็ได้ปล่อย Hadoop ออกสู่สาธารณชนในฐานะ Open Source Project ต่อมา Hadoop จึงตกอยู่ภายใต้การดูแลขององค์กรที่ไม่แสวงหาผลกำไรอย่าง Apache Software Foundation (ASF) อย่างที่เห็นในปัจจุบัน

Hadoop คือ ซอฟท์แวร์ประเภท Open Source ที่จัดทำขึ้นเพื่อเป็นแพลตฟอร์มในการจัดเก็บข้อมูล ซึ่งมีกรอบการทำงานเพื่อใช้ในการจัดเก็บข้อมูลและประมวลผลข้อมูลที่มีขนาดใหญ่มากๆ ที่เราเรียกกันว่า Big Data ซึ่งเจ้าตัว Hadoop เนี่ยก็สามารถปรับขยาย ยืดหยุ่น เพื่อรองรับข้อมูลที่มีจำนวนมากมายมหาศาลได้ ทั้งนี้ก็เพราะมันมีกระบวนการประมวลผลที่แข็งแกร่งมากซึ่งเป็นผลมาจากการประมวลผลข้อมูลแบบกระจายผ่านเครื่องคอมพิวเตอร์ที่ถูกจัดอยู่ในรูปแบบ Cluster อันนำไปสู่ความสามารถในการรองรับข้อมูลที่ไม่จำกัดแถมยังมีความน่าเชื่อถือสูงอีกด้วย

ข้อดี Hadoop

(1) ความสามารถในการรองรับการจัดเก็บข้อมูลขนาดใหญ่มากๆทุกประเภทได้อย่างรวดเร็ว – ด้วยปริมาณข้อมูลในปัจจุบันที่เพิ่มขึ้นอย่างต่อเนื่อง โดยเฉพาะจากแหล่งอย่างโซเชี่ยลมีเดีย และ Internet of Things (IoT) คุณสมบัติข้อนี้ของ Hadoop จึงสำคัญมาก

(2) พลังแห่งการประมวลผล – ด้วยรูปแบบการประมวลผลที่รวดเร็วจากการทำงานแบบ Cluster จึงทำให้ Hadoop กลายเป็นแพลตฟอร์มที่เป็นที่นิยมอย่างกว้างขวางในปัจจุบัน

(3) มีระบบรองรับความผิดพลาด – ด้วยการทำงานแบบ Cluster เมื่อ node ใด node หนึ่งพังลง งานที่มีการทำอยู่ในระบบจะถูกส่งไปยัง node อื่นทันทีเพื่อให้เกิดความต่อเนื่อง รวมถึงระบบเองยังมีการทำข้อมูลสำรองเก็บไว้แบบอัตโนมัติหลายชุดอีกด้วย

(4) ความยืดหยุ่นในการใช้งาน – Hadoop แตกต่างจากระบบฐานข้อมูลดั้งเดิม ที่ต้องมีการแยกประเภทของข้อมูลคร่าวๆก่อนการจัดเก็บ สำหรับ Hadoop เราจะเก็บข้อมูลประเภทไหนก็ได้ มากเท่าไหร่ก็ได้ทันที โดยไม่ต้องมีการแยกประเภทล่วงหน้าแถมยังสามารถเลือกได้อีกว่าจะเอาไปใช้งานด้านใด

(5) ต้นทุนต่ำ – เพราะเป็นแพลตฟอร์มแบบ Open Source จึงสามารถนำมาใช้งานได้ฟรี!

(6) ความสามารถในการขยายการรองรับข้อมูลได้ไม่สิ้นสุด – แค่เพิ่ม node เข้าไปก็สามารถรองรับการจัดเก็บข้อมูลไปได้เรื่อยๆตามแต่เราต้องการ

หลักการเราทราบแล้วว่า Big data คืออะไร ใช้ทำอะไร ทีนี้เรามาดูกันต่อว่าและ Hadoop คืออะไร เกี่ยวข้องอย่างไรกับ Big data ตามหลักการของ Hadoop คือ Java programming framework ที่รองรับการทำงานที่ต้องประมวลผลและเก็บข้อมูลขนาดใหญ่ เป็นส่วนหนึ่งของ apache project โดย Hadoop ถูกออกแบบมาให้เป็น application ที่สามารถทำงานได้บนระบบแบบ node หรือมี hardware จำนวนหลายๆเครื่องพร้อมกัน เพื่อรองรับข้อมูลขนาดใหญ่ ซึ่งใช้การแตก file system ออกมากระจายตาม node ให้สามารถทำงานได้รวดเร็วพร้อมทั้งส่งข้อมูลหากันระหว่าง node ทั้งหมด รวมถึงมีความสามารถใน

การจัดการ node มีเสียหายได้โดนไม่ทำให้เกิดข้อมูลสูญหาย ปัจจุบัน Hadoop ถูกนำมาใช้ในงาน big data จำพวก การคำนวณข้อมูลทางวิทยาศาสตร์เฉพาะทาง, เชิงธุรกิจ รวมถึงวางแผนการขาย และ ประมวลผลข้อมูล sensor จำนวนมาก หรือ internet of things (IOT)

Hadoop ถูกสร้างขึ้นโดยนาย Doug Cutting และ Mike Cafarella ในปี 2006 ซึ่งทำงานที่ Nutch search engine โดยเค้าได้ idea มาจาก Google's Mapreduce ซึ่งเป็น software framework ที่ทำการแตก application ออกเป็นส่วนเล็กๆจำนวนมาก เรียกว่า fragment หรือ block ที่สามารถทำงานบน node แต่ละตัว โดยรวมกันเป็น cluster และทำให้เกิด Hadoop 1.0 ขึ้นมาช่วง พฤศจิกายน 2012 โดยเป็นส่วนหนึ่งใน Apache project

ตั้งแต่ที่เริ่มประกาศ software ออกไป Hadoop ก็ยังถูกพัฒนาอย่างต่อเนื่องจนมาถึง version Hadoop 2 ซึ่งพัฒนาในเรื่องการจัดการทรัพยากรได้ดีขึ้น โดยบริษัทต่างๆสามารถนำเอา Hadoop มาใช้งานใน data center ได้ สามารถขยายเพื่อได้ในรูปแบบ cloud service อย่างเช่น Amazon Web services (AWS), Google Cloud Platform และ Microsoft Azure ซึ่งส่วนมารองรัย Hadoop กันหมดแล้ว

ส่วนประกอบของ Hadoop

อย่างที่บอกว่าเป็น Software framework ทำให้ Hadoop เองประกอบไปดู module จำนวนมาก ซึ่งขั้นต่ำที่จะต้องมีคือ

Hadoop Common: เป็น library ที่จำเป็นในการใช้งานเปรียบได้กับ kernel

Hadoop Distributed File System (HDFS): สำหรับเก็บข้อมูลภายใต้ server node จำนวนมาก ซึ่งต้องอาศัย bandwidth ที่สูงมากในการทำงาน

Hadoop Yet Another Resource Negotiator (YARN): ทำหน้าที่จัดการทรัพยากรที่มีอยู่ รวมถึงรอบการทำงานของ application

Hadoop MapReduce: เป็นรูปแบบ program ที่จัดการกับ ข้อมูลขนาดใหญ่เพื่อสร้างความสัมพันธ์ระหว่างกัน

Big-data-hadoop-apache

Hadoop เองยังสามารถขยายระบบเพิ่มเพื่อรองรับการทำงานที่สูงขึ้น โดยอาศัย software package เหล่านี้

Apache Flume: เครื่องมือที่ใช้เก็บ จัดเรียง และส่งข้อมูบให้กับ HDFS

Apache HBase: เป็น database แบบ nonrelational

Apache Hive: คือ data warehouse ที่ทำการเก็บข้อมูลเพื่อ query และ วิเคราะห์ข้อมูลที่ต้องการ

Cloudera Impala: software ส่วนที่จัดการทำงานแบบ parallel ให้บน Hadoop

Apache Oozie: ไว้จัดการ workflow รวมถึงตั้ง job การทำงานบน Hadoop

Apache Phoenix: ใช้สำหรับ connection ไปยัง HBase และใช้งาน SQL query

Apache Pig: platform ที่ใช้สร้าง program เพื่อทำงานบน Hadoop

Apache Sqoop: เครื่องมือสำหรับส่งข้อมูลขนาดใหญ่ระหว่าง Hadoop และส่วนของข้อมูลที่มีโครงสร้าง หรือ relational database

Apache Spark: engine สำหรับ big data ที่ใช้ประมวลผลข้อมูลแบบ streaming และยังรองรับ SQL ด้วย

Apache Storm: เป็นตัวส่งต่อข้อมูลแบบ real-time

Apache ZooKeeper: เหมือนเป็น server ที่ใช้เก็บข้อมูลทีจะถูกส่งต่อให้อีกที

3)หลักการทำงานของ Hadoop รุ่น 2

องค์ประกอบหลักของ Hadoop

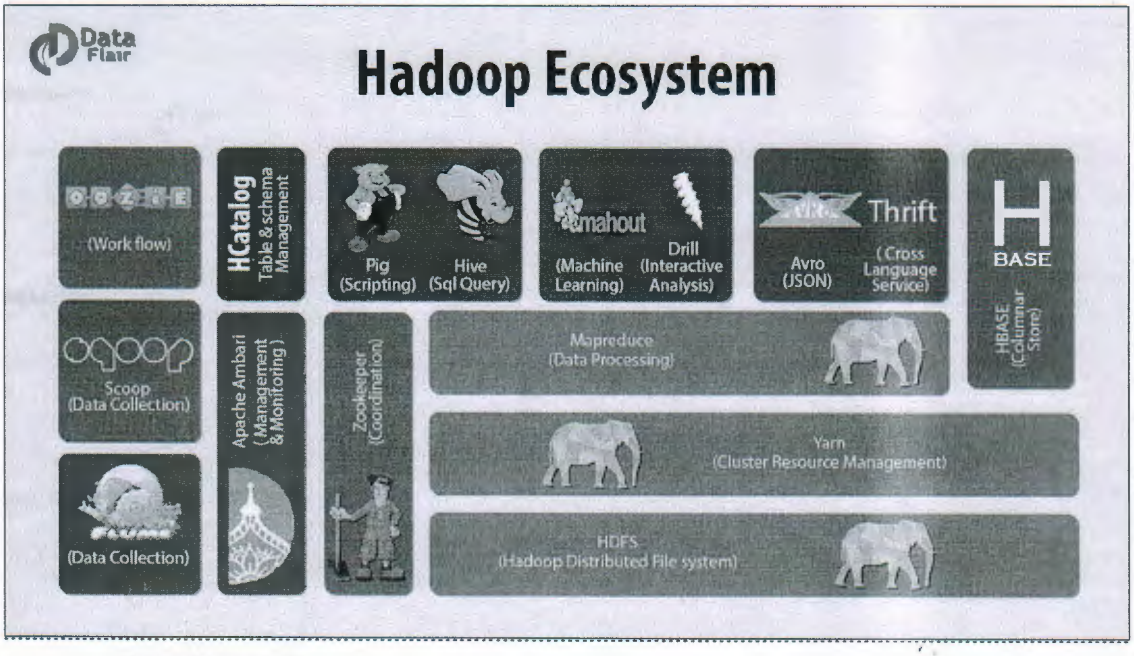
Hadoop ประกอบไปด้วยส่วนประกอบหลักที่สำคัญอยู่ 4 ส่วน ได้แก่

Hadoop Common เป็น libraries และ utilities ส่วนกลางที่ช่วยสนับสนุนการทำงานของ modules อื่นๆ ใน Hadoop หรือบางครั้งเรียกว่า Hadoop ทั่วไป ด้วยตัว Hadoop Common เป็นกลุ่มข้อมูลของคลาส (Class) หรือไรเบอรรี่ (Liberies) จำนวนมาก เพื่อการรองรับการทำงานของ Hadoop มากมาย เช่น บางงานอาจประมวลผลเกี่ยวกับภาพ บางงานอาจประมวลผลข้อความ ดังนั้น Hadoop Common จึงมีในส่วน software จำพวกไรเบอรรี่ (Liberies) เอาไว้เตรียมรับการทำงานทุก ๆ รูปแบบ ยังสามรถปรับเปลี่ยนค่าได้ตรงตามงาน ตามที่จัดเก็บนั้น ๆได้

Hadoop Distributed File System (HDFS) เป็นระบบไฟล์แบบกระจายที่ช่วยทำให้ผู้ใช้สามารถจัดการกับไฟล์ขนาดใหญ่ได้อย่างง่ายดาย นำข้อมูลขนาดใหญ่ มาทำการแบ่งให้เป็นส่วนเล็ก ๆ เรียกว่า Data Splitting นำข้อมูลเล็ก ๆเหล่านี้ กระจายไปประมวลผลยังเครื่องคอมพิวเตอร์ลูกข่าย โดยการทำงานหลัก Hadoopประกอบด้วย 2 ส่วน ส่วนที่ 1 ตัว Master หรือตัวแม่ โดยปกติ ในรุ่นที่ 1 จะมีเพียงตัวเดียว ความสามารถในรุ่นที่ 2 มี Master ได้หลายตัว และอีกส่วนหนึ่งเรียกว่า Slave Node หรือตัวส่วนของลูกอาจมีหลาย ๆ ตัวก็ได้ เอามาช่วยในการประมวลผล เครื่องแม่ข่ายกับเครื่องลูกข่ายสามารถเชื่อมโยงกันโดยผ่านทางเน็ตเวิร์ก หรือผ่านในตัวอินเทอร์เน็ตก็ได้ ยิ่งในปัจจุบันเรามีในส่วนของ Core Computing เอามาช่วยรองรับการประมวลผล

Hadoop YARN (Yet Another Resource Negotiator) ในรุ่นที่ 2 บางครั้งเรียก MRV 2 คือการรองรับการทำงานของ MR จาก8e Mape Reduce ทำให้ที่ในการบริหาร หรือจัดทรัพยากร (Negotiator) Hadoop YARN ยังรองรับการทำงานแบบ Real Time ทำงานให้ถูกต้องเป็น Framework ที่ใช้จัดการ Job scheduling และบริหารจัดการทรัพยากรต่าง ๆ บนระบบ Hadoop cluster

Hadoop MapReduce MapReduce เป็น programming model หรือ programming paradigm ที่ออกแบบมาเพื่อเขียนโปรแกรมสำหรับการประมวลผลแบบขนาน พูดง่าย ๆ คือ เราสามารถเขียนโปรแกรมเพื่อให้มันทำงานบนเครื่องคอมพิวเตอร์หลาย ๆ เครื่องพร้อมกัน ในระบบ Hadoop cluster ของเรา ทำให้การประมวลผลเร็วขึ้นอย่างมาก เช่น การแบ่งหนังสือ เพื่อนับคำ



4) Hadoop HIVE

Hadoop Hive : เป็นเครื่องมือใช้เตรียมข้อมูลที่เป็นลักษณะคลังข้อมูล (Data warehouse) บน Hadoop โดยมีการกำหนด Schema เตรียมไว้ทำให้สามารถทำการสืบค้น (Query) เช่น ชื่อ ที่อยู่ อาชีพ เป็น

ต้น โดยใช้ภาษาที่เรียก Hive QL ซึ่งมีลักษณะคล้ายภาษา SQL จากข้อมูลที่มีการจัดเก็บใน HDFS โดยที่เราไม่ต้องเขียน Map/Reduce เอง เนื่องจาก Hive จะทำการแปลง Hive QL เป็น Map/Reduce แล้วทำการ Execute เป็นแบบ Batch นั่นเอง และสามารถ Access ด้วย ODBC/JDBC ตัวอย่างหน้าจอ Hive ของค่าย Cloudera แนะนำ Apache Hive เป็น Open Sources เพราะในโลกการประมวลผลข้อมูลโดยทั่วไปที่ต้องการพึ่งพาอาศัยประสิทธิภาพของ Hadoop ที่ต้องความเร็วในการประมวลข้อมูลปริมาณมาก ด้วยการย้าย Operation Data เป็นปริมาณมากๆในระดับ 100 ล้านรายการมาเก็บใน Hadoop เพื่อการประมวลผลแบบสรุปข้อมูลเชิงสถิติแล้วส่งคืนกลับไปให้ที่เก็บข้อมูลหลักของ Operation Data ให้ใช้ในเงื่อนไขการตัดสินใจของระบบซอฟต์แวร์หลักที่ประมวลผลทางธุรกิจ หรือ ส่งต่อไปให้ระบบข้อมูลเพื่อการแสดงผลแบบ Dashboard ของ BI (Business Intelligence)หรือ การย้าย Operation Data จากถังเก็บข้อมูลปกติในระบบปฏิบัติงานออกไปเก็บที่ Hadoop เพื่อทำให้จำนวนข้อมูลใน Operation Data มีปริมาณน้อยลงที่จะส่งผลให้การเรียกข้อมูลมาปฏิบัติงานได้เร็วขึ้นด้วยการย้ายส่วนเรียกข้อมูลเก่ากว่าเวลาปกติ หรือ มีปริมาณมากให้มาเรียกดูข้อมูลที่ย้ายมาไว้บน Hadoopแทน เมื่อข้อมูลที่ต้องการถูกนำมาวางไว้ที่ Hadoop การเรียกข้อมูลแบบ Random Access หรือ เข้าถึงตัวข้อมูลที่ต้องการได้ทันทีด้วยวิธีเดียวกับการใช้ภาษา SQL แบบปกติที่ใช้ในการจัดการข้อมูลแบบตารางฐานข้อมูลโดยทั่วไปยังเป็นความต้องการหนึ่งที่จะทำให้การปฏิบัติงานกับข้อมูลยังเป็นวิธีปกติเหมือนที่เคยปฏิบัติมา

Apache Hive เป็นบริการหนึ่งในครอบครัวของ Hadoop Ecosystem ที่ทำงานอยู่บนพื้นฐานของ HDFS ของ Hadoop ที่รองรับการการนำเสนอข้อมูลในรูปแบบโครงสร้างตารางข้อมูลที่สามารถใช้ภาษา SQL ในการเรียกข้อมูลตามปกติได้ และ มีคำสั่งที่สามารถโอนข้อมูลจากตารางข้อมูลในฐานข้อมูลของ Operation Data ที่เป็น RDBMS รวมถึงมี API ที่สมบูรณ์สำหรับการพัฒนาโปรแกรมที่เรียกข้อมูลไปประมวลเฉพาะแอพพลิเคชั่นที่พัฒนาเพื่อวัตถุประสงค์การใช้งานทั่วไปด้วย Apache Hive ที่ทำงานอยู่บน Hadoop เชิงเทคนิคกันสักเล็กน้อย โดยกระบวนการทำงานหลัก ๆ นั้น Hive ทำหน้าที่ส่งคำสั่งประมวลผลข้อมูลผ่านเข้าสู่ Hadoop ซึ่งการประมวลผลใดๆกับข้อมูล ของ Hadoop ยังคงทำผ่าน MapReduce ตามกระบวนท่าปกติของ Hadoop ดังนั้นบทบาทของ Hive จะเป็นตัวกลางที่มี Engine ในการแปลงภาษา HQL หรือ SQL ในแบบของ Hive เป็น MapReduce เพื่อส่งไปทำงานที่ Hadoop Engine โดยกระบวนทำงานของ Hive Engine จำเป็นต้องมี Metadata ที่เป็นข้อมูลอธิบายรายละเอียดของโครงสร้างเสมือนตารางข้อมูล ประเภทข้อมูลที่อ้างอิงเป็นคอลัมน์ข้อมูล ซึ่งภาพของการจัดการกับ Metadata คือ การสร้าง Table หรือ Create Table แบบภาษา SQL นั่นเอง การที่ต้องมี Metadata ของ Table ก็เพื่อให้ Hive Engine และ Hadoop Engine ทำงานร่วมกัน โดยหลักการทำงานคือการแปลง HQL เป็น MapReduce ในการประมวลผลข้อมูล หรือ แปลงผลลัพธ์การประมวลผลข้อมูลจาก MapReduce ที่อ่านจากไฟล์คืนกลับเป็นข้อมูลแบบมีโครงสร้างแบบตารางข้อมูลที่เอื้อต่อการมี API ของแต่ละภาษาโปรแกรมมิ่งมาต่อยอดนำไปสู่การทำวิเคราะห์ข้อมูลประมวลผลข้อมูลได้อย่างง่ายๆ แสดงตัวอย่างกระบวนการทำงานตามรูปที่ 1 ดังนั้นส่วนประกอบที่สำคัญของ Hive ที่ Hive จะสร้าง MapReduce Job ตอนที่เราทำการบันทึกข้อมูล หรือ เรียกดูข้อมูล คือตัว Metastore

5) Apache pig

Apache Pig ภาษาสคริปต์ เป็นนามธรรมเหนือ MapReduce เป็นเครื่องมือ / แพลตฟอร์มที่ใช้ในการวิเคราะห์ชุดข้อมูลขนาดใหญ่ที่แสดงเป็นกระแสข้อมูล โดยทั่วไปจะใช้หมูกับ Hadoop: เราสามารถดำเนินการจัดการข้อมูลทั้งหมดใน Hadoop โดยใช้ Apache Pig ในการเขียนโปรแกรมวิเคราะห์ข้อมูล Pig ให้ภาษาระดับสูงที่เรียกว่า Pig Latin. ภาษานี้จัดเตรียมตัวดำเนินการต่างๆโดยใช้ซึ่งโปรแกรมเมอร์สามารถพัฒนาฟังก์ชันของตนเองสำหรับการอ่านเขียนและประมวลผลข้อมูล เพื่อวิเคราะห์ข้อมูลโดยใช้ Apache Pig โปรแกรมเมอร์ต้องเขียนสคริปต์โดยใช้ภาษา Pig Latin สคริปต์ทั้งหมดเหล่านี้ถูกแปลงเป็นแผนที่และลดงานภายใน Apache Pig มีส่วนประกอบที่เรียกว่า Pig Engine ที่ยอมรับสคริปต์ Pig Latin เป็นอินพุตและแปลงสคริปต์เหล่านั้นเป็นงาน MapReduce หากโปรแกรมเมอร์ที่ไม่ค่อยเก่ง Java มักใช้ในการต่อสู้กับ Hadoop

โดยเฉพาะอย่างยิ่งในขณะที่ทำงาน MapReduce ใด ๆ Apache Pig เป็นประโยชน์สำหรับโปรแกรมเมอร์ทุกคน การใช้ Pig Latinโปรแกรมเมอร์สามารถทำงาน MapReduce ได้อย่างง่ายดายโดยไม่ต้องพิมพ์โค้ดที่ซับซ้อนใน Java Apache Pig ใช้ multi-query approachซึ่งจะช่วยลดความยาวของรหัส ตัวอย่างเช่นการดำเนินการที่ต้องการให้คุณพิมพ์รหัส 200 บรรทัด (LoC) ใน Java สามารถทำได้อย่างง่ายดายโดยพิมพ์น้อยเพียง 10 LoC ใน Apache Pig ในที่สุด Apache Pig จะลดเวลาในการพัฒนาลงเกือบ 16 เท่า Pig Latin คือ SQL-like language และเรียนรู้ Apache Pig ได้ง่ายเมื่อคุณคุ้นเคยกับ SQL

Apache Pig มีตัวดำเนินการในตัวจำนวนมากเพื่อรองรับการดำเนินการกับข้อมูลเช่นการรวมตัวกรองการสั่งซื้อและอื่น ๆ นอกจากนี้ยังมีประเภทข้อมูลที่ซ้อนกันเช่นสิ่งที่เพิ่มขึ้นถุงและแผนที่ที่ไม่มีใน MapReduce เหมาะกับเหมาะกับการทำ ETL สำหรับการแปลงข้อมูลในรูปแบบต่าง ๆ เช่น JSON

คุณสมบัติของหมู

Apache Pig มาพร้อมกับคุณสมบัติดังต่อไปนี้ -

Rich set of operators - มีตัวดำเนินการจำนวนมากในการดำเนินการเช่น join, sort, filer ฯลฯ

Ease of programming - Pig Latin คล้ายกับ SQL และง่ายต่อการเขียนสคริปต์ Pig ถ้าคุณเก่ง SQL

Optimization opportunities - งานใน Apache Pig จะเพิ่มประสิทธิภาพการดำเนินการโดยอัตโนมัติดังนั้นโปรแกรมเมอร์จึงต้องเน้นเฉพาะความหมายของภาษาเท่านั้น

Extensibility - การใช้ตัวดำเนินการที่มีอยู่ผู้ใช้สามารถพัฒนาฟังก์ชันของตนเองเพื่ออ่านประมวลผลและเขียนข้อมูลได้

UDF's - หมูให้สิ่งอำนวยความสะดวกในการสร้าง User-defined Functions ในภาษาโปรแกรมอื่น ๆ เช่น Java และเรียกใช้หรือฝังไว้ใน Pig Scripts

Handles all kinds of data- Apache Pig วิเคราะห์ข้อมูลทุกประเภททั้งแบบมีโครงสร้างและแบบไม่มีโครงสร้าง จะจัดเก็บผลลัพธ์ใน HDFS

การใช้งาน Apache Pig

โดยทั่วไปนักวิทยาศาสตร์ด้านข้อมูลจะใช้ Apache Pig ในการปฏิบัติงานที่เกี่ยวข้องกับการประมวลผลเฉพาะกิจและการสร้างต้นแบบอย่างรวดเร็ว ใช้ Apache Pig -

-เพื่อประมวลผลแหล่งข้อมูลขนาดใหญ่เช่นบันทึกเว็บ

-เพื่อดำเนินการประมวลผลข้อมูลสำหรับแพลตฟอร์มการค้นหา

ในการประมวลผลการโหลดข้อมูลที่อ่อนไหวต่อเวลา (log file)

6) Apache Sqoop

เป็นเครื่องมือที่ทำหน้าที่ในการ Transfer ข้อมูลจากระบบฐานข้อมูลอื่น ๆ เช่น Oracle, SQL Server, My SQL เข้ามาเก็บในรูปแบบ HDFS ของ Hadoop (ซึ่งเราต้องสร้าง Connection ผ่าน jdbc เพื่อต่อไปยังระบบฐานข้อมูล และสร้าง Link เพื่อเชื่อมต่อไปยัง HDFS) จากข้อมูลมีโครงสร้าง (Structured Data) เป็นแบบไม่มีโครงสร้าง (Unstructured Data) หรือกึ่งมีโครงสร้าง ข้อมูลกึ่งมีโครงสร้าง (Semi-Structured Data) ประยุกต์ใช้ ระบบธนาคาร ระบบสายการบิน

7) Apache Mahout

เป็นเครื่องมือของ Data Scientist ที่ต้องการทำPredictive Analytics ข้อมูลบน Hadoop โดยใช้ภาษาจาวา ทั้งนี้ Mahout สามารถใช้ Algorithm ที่เป็น Recommender, Classification และ Clustering ได้ เป็นเครื่องมือ Open Source ที่เอาไว้ใช้ทำ machine learning algorithms จากข้อมูลขนาดใหญ่บน Hadoop เช่น Recommendation, Classification, Clusteringเช่น บัตรเครดิต การ์คัดลอดวรรณกรรม ตรวจสอบลิขสิทธิ การคัดกรองอีเมล์ การพยากรณ์อากาศ.....

8) Apache Zookeeper

เหมือนเป็น server ที่ใช้เก็บข้อมูลที่จะถูกส่งต่อให้อีกที Zookeeper เป็นระบบที่สร้างขึ้นตามแนวคิดของ Chubby lock service ที่สร้างขึ้นโดย Google เช่นเดียวกันกับ Hadoop ที่สร้างขึ้นจาก MapReduce หน้าที่ของ Zookeeper คือเป็นระบบที่ใช้ทำกระบวนการ Distributed Coordination ซึ่งจำเป็นสำหรับการประมวลผลแบบกระจายศูนย์ เช่น Leader Election (เลือกโพรเซสที่เป็นผู้นำ) Distributed Locking (ล็อคทรัพยากรเพื่อประมวลผล) ซึ่งเป็นสิ่งจำเป็นต่อการสร้างระบบที่อาศัยการทำงานร่วมกันของคอมพิวเตอร์หลายเครื่อง แม้แต่ระบบเช่น Hadoop หรือ HBase ก็พึ่งพา Zookeeper เช่นกัน ทำซ้ำข้อมูล (data replication) ของคอมพิวเตอร์หลายเครื่อง Zookeeper ลบข้อมูลที่ซ้ำมากเกินไป..... .....

9) ฮอนทอนเวิร์ก (Hortonworks)

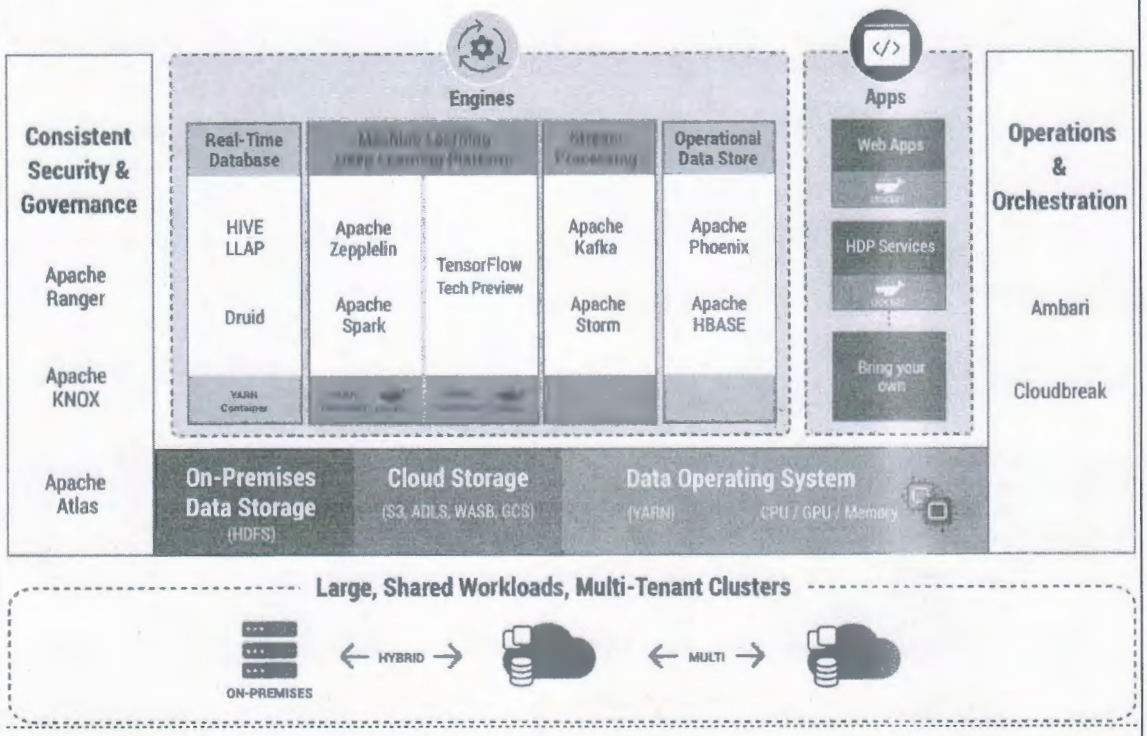
Hortonworks ที่สร้างมาจากพื้นฐานของ Open Source Software Solution สำหรับ Big Data Analytics โดยเฉพาะ Hortonworks ถือเป็นผู้ผลิตรายใหญ่รายหนึ่งที่ออกมาตอบโจทย์ความต้องการข้างต้นให้กับเหล่าองค์กรโดยเฉพาะ ด้วยการพัฒนา Distribution ของระบบ Big Data Analytics ที่ใช้ Open Source Software ซึ่งผ่านการทดสอบ แก้ไข และควบคุมคุณภาพมาเรียบร้อยแล้ว พร้อมทั้งเสริมความสามารถให้การติดตั้งใช้งานหรือดูแลรักษาทำได้อย่างง่ายดายยิ่งขึ้น พร้อมให้เหล่าองค์กรนำไปใช้ได้โดยไร้กังวล โซลูชันของ Hortonworks นำเอาโมดูลพื้นฐานของ Open Source Software ที่โดดเด่นมาประกอบขึ้นเป็นแพลตฟอร์มเวอร์ชันที่พร้อมใช้งานสำหรับองค์กร ด้วยกัน 2 โซลูชัน ดังนี้

Hortonworks Data Platform (HDP) สำหรับสร้าง Big Data Analytics อย่างมีประสิทธิภาพ HDP นี้เป็นพัฒนามาจาก Apache Hadoop Distribution สำหรับองค์กรโดยเฉพาะ ด้วยการใช้สถาปัตยกรรมของ YARN เป็นหลัก โดยมีวัตถุประสงค์เพื่อทำการจัดเก็บและวิเคราะห์ข้อมูลขนาดใหญ่จำนวนมหาศาล โดยองค์กรสามารถรวบรวมข้อมูลจากระบบต่าง ๆ ให้ผู้ใช้งานสามารถนำข้อมูลเหล่านี้ไปใช้วิเคราะห์ตอบโจทย์และต่อยอดได้หลากหลาย นอกจากความสามารถในการจัดการข้อมูลขนาดใหญ่และการวิเคราะห์ข้อมูลอันชาญฉลาดแล้ว HDP ยังรวมเอาการจัดการด้าน Security และ Governance เพื่อตอบโจทย์สำหรับองค์กรที่อาจมีการนำข้อมูลสำคัญทางธุรกิจ มาใช้วิเคราะห์และมีความปลอดภัยในการเข้าถึง ไปจนถึงมีระบบในการบริหารจัดการแบบรวมศูนย์หรือ Orchestration เพื่อให้การจัดการ Big Data Analytics เป็นไปได้อย่างอัตโนมัติและมีประสิทธิภาพในเชิงการวิเคราะห์ข้อมูลนั้น HDP สามารถรองรับได้ทั้งแต่ระดับของการ Query ข้อมูลทั่วไป, การแปลงข้อมูลรูปแบบต่าง ๆ ให้สามารถเชื่อมต่อและเข้าถึงโดยวิธีการที่กำหนด, การนำข้อมูลจากหลายแหล่งมาวิเคราะห์ร่วมกัน, การทำ Machine Learning ไปจนถึงการทำ Deep Learning เพื่อต่อยอดไปยังระบบ AI ได้ทันที

HDP สามารถใช้งานได้ทั้งแบบ On-premises, Public Cloud, Hybrid Cloud ไปจนถึง Multi-Tenant Cluster ทำให้สามารถตอบโจทย์ได้ทั้งสำหรับธุรกิจขนาดกลางไปจนถึงขนาดใหญ่ได้อย่างครบถ้วน Hortonworks Data Flow (HDF) จัดการ Real-time and Live Stream Data ปริมาณมหาศาล รองรับ IoT ถึงแม้ความสามารถของ HDP นั้นจะถือว่าหลากหลายและตอบโจทย์ของธุรกิจในส่วนของ Big Data Analytics ได้แล้ว แต่ HDP เองนั้นก็ไม่ได้ครอบคลุมการใช้งานได้ในทุกกรณี เนื่องจากการวิเคราะห์ข้อมูลที่ถูกส่งเข้ามาแบบ Real-time นั้นต้องมีการใช้เทคโนโลยีที่แตกต่างออกไป และ Hortonworks ก็ให้ความสำคัญกับการใช้งานในส่วนนี้ โดยทำการพัฒนาผลิตภัณฑ์อย่าง HDF ขึ้นมาเสริมความสามารถด้านการวิเคราะห์ Real-time Data โดยเฉพาะ เพื่อรองรับ Internet of Things (IoT) Application ที่กำลังเติบโตอย่างรวดเร็วในปัจจุบัน

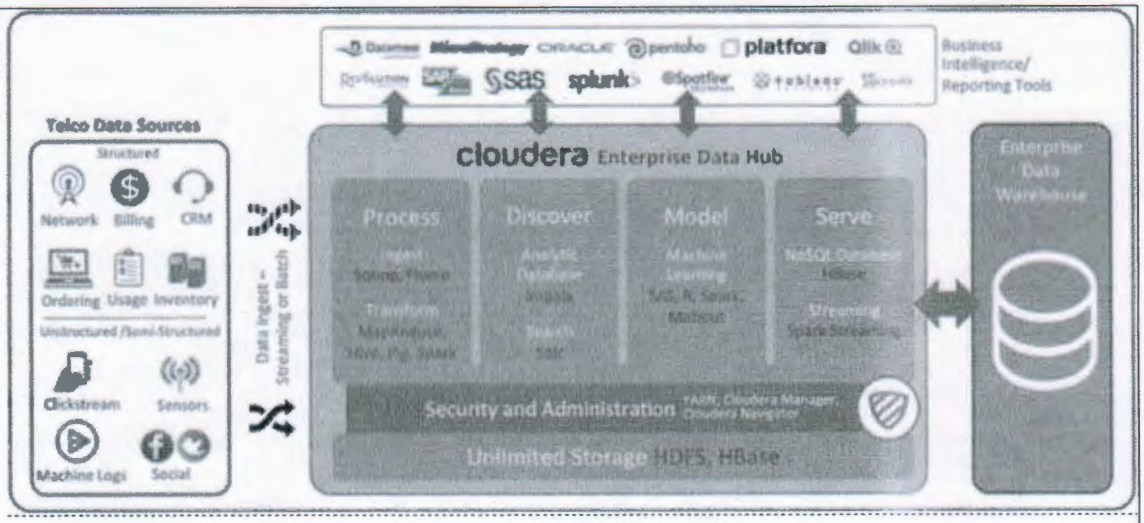
HDF นี้จะทำการรวมเอาเทคโนโลยี Open Source Software ทางด้าน Flow Management และ Streaming Analytics อย่าง Apache NiFi, MiNiFi, Kafka, และ Storm เข้าไว้ด้วยกัน ทำให้สามารถวิเคราะห์ข้อมูลที่ถูกส่งเข้ามาเป็นแบบ Live Stream จากแหล่งต่าง ๆ ได้อย่างทันท่วงที อีกทั้งยังสามารถทำ

การส่งข้อมูลเชื่อมต่อไปยัง HDP เพื่อทำการจัดเก็บข้อมูลเอาไว้ใช้ในการวิเคราะห์เพิ่มเติมภายหลังได้อีกด้วย.....



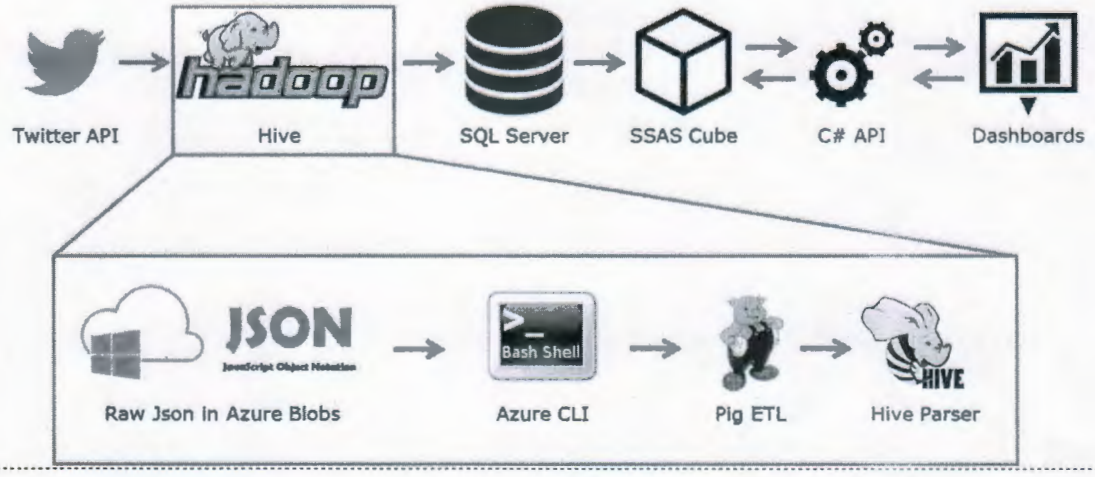
10) คลาวด์เดอรา (Cloudera)

Cloudera เป็นแพลตฟอร์มซอฟต์แวร์สำหรับการจัดการ และวิเคราะห์ข้อมูลที่มีจำนวนมาก ถือเป็น Hadoop ที่ช่วยให้การจัดการ และวิเคราะห์ข้อมูลที่รวดเร็ว ง่าย และปลอดภัย ทำให้องค์กรสามารถใช้ประโยชน์จากข้อมูลได้มากที่สุด โดยล่าสุด ได้ทำการเปิดตัว Cloudera Enterprise 5.7 ที่มีประสิทธิภาพในการประมวลผลข้อมูลสูงขึ้น 3 เท่า และมีความเร็วในการวิเคราะห์ข้อมูลเพิ่มเป็น 2 เท่า โซลูชัน Cloudera Enterprise มาใช้ในการสร้างประสบการณ์การใช้งานของลูกค้า ทำให้สามารถคัดสรรบริการที่เหมาะสมสำหรับแต่ละบุคคล แถมยังช่วยลดอัตราการยกเลิกการใช้บริการ และช่วยในการเสริมให้เกิดรายได้ใหม่ๆ เข้ามา โดยได้ใช้กับ 5 กลุ่มธุรกิจประกอบด้วย True Online, True Move, True Visions, True Money และ True Life การเข้ามาของ Big Data จึงได้ทำการตัดสินใจเลือกติดตั้งโซลูชันที่มีความยืดหยุ่น และอยู่บนมาตรฐานระบบเปิด ที่มีทั้งความง่าย และยังสามารถนำเข้ามาเสริมในเรื่องการพัฒนาเทคโนโลยีที่ เกี่ยวกับเรื่องของบิ๊กดาต้าได้อีกด้วย โดย Cloudera นั้นได้นำเสนอผลิตภัณฑ์ที่ทำให้เทคโนโลยีอย่าง Hadoop นั้นในแง่ของการนำมาใช้ และบริหารจัดการกลายเป็นเรื่องง่าย และรวดเร็ว รวมถึงมีความปลอดภัยด้วยCloudera ทำงานร่วมกับบริษัทบริการทางการเงินระดับโลกกว่า 180 องค์กร อันได้แก่ Bank Mandiri , Credit Suisse , DBS Bank , Nordea , Northern Trust และ Royal Scotland เป็นต้น Cloudera ช่วยให้บริษัทชั้นนำเหล่านี้สามารถ สร้างความเข้าใจเชิงลึกจากข้อมูลของลูกค้า (deeper customer insights) สามารถหาจุดบกพร่องในการทำงานที่ผ่านมาและสร้างความปลอดภัยในการทำธุรกรรมของลูกค้าให้สูงขึ้นและลดต้นทุนในการปฏิบัติตามกฎระเบียบของทางการ.....



11) แซส (SAS)

"แซส" (SAS) ผู้นำตลาดซอฟท์แวร์และบริการด้านการวิเคราะห์ข้อมูลเชิงธุรกิจ (Business Analytics) ด้วยโซลูชั่นส์เชิงนวัตกรรมทางด้านการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics) สามารถช่วยเพิ่มประสิทธิภาพในการตัดสินใจของลูกค้าผ่านข้อมูลที่นำมาวิเคราะห์อย่างละเอียดเพื่อผลตอบรับที่ดีที่สุดในเวลาที่สั้นที่สุด ทั้งนี้ได้ตระหนักถึงคุณประโยชน์นานับประการของระบบฮาดูป (Hadoop) รวมถึงอัตราการเติบโตของการใช้งานในระบบฮาดูปทั่วโลกที่เพิ่มสูงขึ้นอย่างต่อเนื่อง จากการที่ระบบฮาดูปสามารถลดการใช้พื้นที่ในการจัดเก็บข้อมูลขนาดใหญ่และสามารถเรียกข้อมูลเพื่อนำมาวิเคราะห์ได้อย่างรวดเร็วกว่าระบบการจัดเก็บฐานข้อมูลแบบเดิม จึงทำให้ได้ผลลัพธ์ที่รวดเร็วและมีค่าใช้จ่ายต่ำกว่าระบบอื่นๆ ด้วยเหตุนี้ทางแซสจึงผลิตซอฟท์แวร์โซลูชั่นส์ต่างๆเพื่อรองรับการทำงานบนฮาดูปเพื่อตอบสนองต่อองค์กรธุรกิจที่มีข้อมูลขนาดใหญ่ในทุกอุตสาหกรรม ซึ่งนอกจากจะช่วยด้านบริหารจัดการข้อมูลมหาศาลบนฮาดูปแล้วยังรวมไปถึงการวิเคราะห์ข้อมูลได้ทันทีตลอดจนสามารถแสดงผลวิเคราะห์อย่างแม่นยำรวดเร็ว และง่ายต่อการเข้าใจผ่านอุปกรณ์แอพพลิเคชั่นต่างๆเรียกได้ว่าเป็นการทำงานแบบวัฏจักร (Life Cycle) อย่างสมบูรณ์แบบที่สุดของอนาไลติกส์บนระบบฮาดูปที่มีในขณะนี้.....



12) Apache Hadoop

คือโครงการ Opensource Software สำหรับการสร้างระบบ Distributed Computing ที่มีความเสถียรสูง และสามารถเพิ่มขยายได้อย่างมหาศาล โดยตัวอย่างของผู้ที่ใช้งาน Apache Hadoop ตัว Hadoop เองยังสามารถขยายระบบเพิ่มเพื่อรองรับการทำงานที่สูงขึ้น โดยอาศัย software package เหล่านี้

Apache Flume: เครื่องมือที่ใช้เก็บ จัดเรียง และส่งข้อมูลให้กับ HDFS

Apache HBase: เป็น database แบบ nonrelational

Apache Hive: คือ data warehouse ที่ทำการเก็บข้อมูลเพื่อ query และ วิเคราะห์ข้อมูลที่ต้องการ

Cloudera Impala: software ส่วนที่จัดการทำงานแบบ parallel ให้บน Hadoop

Apache Oozie: ไว้จัดการ workflow รวมถึงตั้ง job การทำงานบน Hadoop

Apache Phoenix: ใช้สำหรับ connection ไปยัง HBase และใช้งาน SQL query

Apache Pig: platform ที่ใช้สร้าง program เพื่อทำงานบน Hadoop

Apache Sqoop: เครื่องมือสำหรับส่งข้อมูลขนาดใหญ่ระหว่าง Hadoop และส่วนของข้อมูลที่มีโครงสร้าง หรือ relational database

Apache Spark: engine สำหรับ big data ที่ใช้ประมวลผลข้อมูลแบบ streaming และยังรองรับ SQL ด้วย

Apache Storm: เป็นตัวส่งต่อข้อมูลแบบ real-time

ตัวอย่างของผู้ที่ใช้งาน Apache Hadoop นั้นมีดังนี้

Facebook

Facebook นั้นมี Apache Hadoop Cluster อยู่ด้วยกัน 2 ชุด ชุดแรกประกอบจาก Server จำนวน 1,100 เครื่อง, CPU 8,800 Cores และพื้นที่ 12PB (12,000TB) และชุดที่สองประกอบจาก Server จำนวน 300 เครื่อง, CPU 2,400 Cores และพื้นที่ 3PB (3,000TB)

Yahoo!

Yahoo! นั้นใช้ Server มากกว่า 40,000 เครื่อง, CPU มากกว่า 100,000 ชุดสำหรับรองรับระบบ Ads และ Web Search

นอกจากนี้ยังมีผู้ให้บริการรายใหญ่ๆ มากมายอย่าง Twitter, ImageShack, Adobe, AOL และอื่นๆ อีกมากมายที่ใช้ Apache Hadoop ในการจัดเก็บข้อมูลแทนฐานข้อมูลแบบ SQL รวมถึง Microsoft เองก็มีแผนที่จะให้ MS SQL สามารถทำงานเชื่อมต่อกับ Apache Hadoop ได้เช่นกัน อีกทั้งผู้ผลิตรายใหญ่ๆ อย่าง IBM และ Supermicro เองก็ให้การสนับสนุน Apache Hadoop กันเป็นอย่างมากอีกด้วยส่วนในวงการการศึกษาและทางภาคธุรกิจเอง Apache Hadoop ถือเป็นทางเลือกที่ดีในงานหลายๆ ประเภท ไม่ว่าจะเป็นงานประมวลผลประสิทธิภาพสูง (High Performance Computing), Scientific Computing, Image Processing, Information Retrieval, Machine Learning, Social Network Analysis, Data Mining, Business Intelligence (BI), Network Security, Sensor Data Storage, Biomedical, Statistic, Machine Translation, Language Modeling, Bioinformatic, Email Analysis และอื่น ๆ อีกมากมาย และแนวโน้มการเติบโตของ Apache Hadoop ก็จะยังคงมีต่อไปเรื่อยๆ อีกด้วย

Hardware สำหรับใช้งาน Apache Hadoop

เนื่องจาก Apache Hadoop นั้นเป็น Opensource ที่สามารถทำงานร่วมกันได้กับทั้ง Linux และ Microsoft Windows โดยเรียกใช้งาน Java เป็นหลัก ดังนั้น Apache Hadoop จึงสามารถทำงานบน Hardware ได้หลากหลาย โดยต้องทำการเลือก Spec สำหรับ Hardware ที่เหมาะสมสำหรับแต่ละหน้าที่ใน Cluster ของ Apache Hadoop ให้ดี ที่อเมริกา Supermicro เป็น Server ที่ได้รับความนิยมอย่างสูงสำหรับทำ Apache Hadoop มาก เนื่องจากมีความหลากหลายของ Hardware ที่สามารถปรับแต่งและเลือกใช้ให้เหมาะสมกับแต่ละหน้าที่ใน Hadoop และแต่ละ Project ที่แตกต่างกันได้ ทาง Supermicro จึงได้จัดชุดของ Hardware ที่เหมาะสำหรับการทำ Apache Hadoop มาดังนี้

Apache Hadoop Name Node – เป็น Server ขนาด 1U สำหรับทำหน้าที่เป็น Apache Hadoop Name Node ติดตั้ง CPU Intel E5600 1 ชุด, หน่วยความจำ 48GB และ Hard Drive แบบ 147GB SAS 15k 2 ชุด

Apache Hadoop Data Node 1U – เป็น Server ขนาด 1U สำหรับทำหน้าที่เป็น Apache Hadoop Data Node ติดตั้ง CPU Intel E5600 1 ชุด, หน่วยความจำ 24GB และ Hard Drive แบบ 2TB SATA จำนวน 4 ชุด รวมพื้นที่ 8TB ต่อ 1U เหมาะสำหรับระบบที่ต้องการลงทุนเริ่มต้นทีละเล็กน้อย

Apache Hadoop Data Node 2U – เป็น Server ขนาด 2U สำหรับทำหน้าที่เป็น Apache Hadoop Data Node ติดตั้ง CPU Intel E5600 1 ชุด, หน่วยความจำ 24GB และ Hard Drive แบบ 2TB SATA จำนวน 12 ชุด รวมพื้นที่ 24TB ต่อ 2U เหมาะสำหรับระบบที่ต้องการใช้พื้นที่ขนาดใหญ่

Apache Hadoop Twin Data Node 2U – เป็น Twin Server ขนาด 2U สำหรับทำหน้าที่เป็น Apache Hadoop Data Node มี Server 2 ชุดแบบ Hot Swap โดยแต่ละชุดติดตั้ง CPU Intel E5600 1 ชุด, หน่วยความจำ 24GB และ Hard Drive แบบ 2TB SATA จำนวน 12 ชุด รวมพื้นที่ 24TB ต่อ 2U เหมาะสำหรับระบบที่ต้องการใช้พื้นที่ขนาดใหญ่ และต้องการหน่วยประมวลผลจำนวนมาก

1GbE / 10GbE Ethernet Switches – สำหรับเชื่อมต่อระบบ Cluster ทั้งหมดด้วยความเร็ว 1GbE หรือ 10GbE ตามแต่ความต้องการของแต่ละระบบ

IPMI Server Management – เป็นระบบสำหรับบริหารจัดการ Server จากระยะไกล โดยสามารถทำ Remote KVM-over-LAN และ Virtual Media-over-LAN เพื่อให้สามารถบริหารจัดการได้เสมือนทำงานอยู่ที่หน้าเครื่อง Server แต่ละเครื่องได้ผ่านทาง LAN โดยไม่ต้องมี Hardware เพิ่มเติม

ในอนาคต Apache Hadoop และ Big Data Solution จะกลายเป็นสิ่งที่เข้ามามีบทบาทในระดับ Enterprise มากขึ้นเรื่อยๆ โดยค่าใช้จ่ายของ Hardware โดยรวมจะประหยัดกว่าระบบจัดเก็บข้อมูลแบบในปัจจุบัน เนื่องจากการนำ Server มาใช้งานในลักษณะ Cloud ทำให้สามารถตัดค่าใช้จ่ายของระบบจัดเก็บข้อมูลแบบ SAN Storage หรือ NAS Storage ได้ โดยมีความสามารถในการเก็บรักษาข้อมูลในระดับที่สูงขึ้นอีกด้วย ดังนั้นการเลือกใช้ Hardware ให้เหมาะสมในระบบ Cloud จึงกลายเป็นสิ่งสำคัญตามมา ในขณะที่ผู้ดูแลระบบเองก็ควรจะต้องเริ่มศึกษาเทคโนโลยีใหม่ ๆ เพิ่มเติมนอกเหนือจาก SQL Database แบบเดิมๆ ด้วยเช่นกัน.....

2.2 ประสบการณ์/ประโยชน์ที่ได้รับ/การประยุกต์ใช้กับหน่วยงาน

☑ต่อตนเอง

ทำให้สามารถเรียนรู้การรวบรวมข้อมูลและวิเคราะห์ข้อมูล Big data ผ่านโปรแกรม Hadoop โดยใช้โปรแกรมเสริมใน Hadoop.....

.....

☑ต่อหน่วยงาน / การนำมาประยุกต์ใช้กับหน่วยงาน

.....แนวทางการวิเคราะห์ข้อมูล Big data ผ่านโปรแกรม Hadoop ทำพยากรณ์อากาศ การทำ Infographic จาก Data ให้สามารถใช้ dashboard ได้.....

.....

2.3 ปัญหาและอุปสรรคในการอบรม/สัมมนา/พัฒนาความรู้ฯ

.....- การเรียนรู้เน้นด้านการพยากรอากาศ จำเป็นต้องฝึกมากกว่าปกติ ใช้เวลามาก.....

.....

.....

2.4 ข้อคิดเห็นและข้อเสนอแนะ

.....- น่าสามารถประยุกต์ใช้กับพืชขาดธาตุอาหาร และขาดน้ำได้ แต่จำเป็นต้องสร้างฐานข้อมูลใหญ่.....

.....

ลงชื่อ.....

(นายดิเรก คงแพ)

ตำแหน่ง นักวิเคราะห์นโยบายและแผนชำนาญการพิเศษ

ผู้รายงาน

วันที่ ๓๑ เดือน มกราคม พ.ศ. ๒๕๖๕

ส่วนที่ 3 ความเห็นของผู้บังคับบัญชา

(✓) ทราบ

.....

.....

.....

ลงชื่อ.....

(นายสมศักดิ์ สุขจันทร์)

ตำแหน่ง ผู้อำนวยการกองนโยบายและแผนการเชื้อเพลิง

วันที่ ๒ เดือน ก.พ. พ.ศ. ๖๕