การดึงและรวบรวมข้อมูลขนาดใหญ่เพื่อนำมาประมวลผลให้เกิดประโยชน์และมีประสิทธิภาพสูงสุดด้วยเครื่องมือ Big Data

BIG DATA GATHERING FOR THE MOST EFFECTIVE ANALYSIS BY USING BIG DATA TOOLS

นาย ภูชิสส์ วัฒนกรวิโรจน์

โครงงานสหกิจศึกษานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยี ไทย-ญี่ปุ่น
พ.ศ. 2561

การดึงและรวบรวมข้อมูลขนาดใหญ่เพื่อนำมาประมวลผลให้เกิดประโยชน์และมีประสิทธิภาพสูงสุดด้วยเครื่องมือ Big Data

BIG DATA GATHERING FOR THE MOST EFFECTIVE ANALYSIS BY USING BIG DATA TOOLS

นาย ภูชิสส์ วัฒนกรวิโรจน์

โครงงานสหกิจศึกษานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรบัณฑิต สาขาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย - ญี่ปุ่น
ปีการศึกษา 2561

คณะกรรมการสอบ

..... ประธานกรรมการสอบ
(อาจารย์ อมรพันธ์ ชมกลิ่น)

..... กรรมการสอบ
(อาจารย์ โอพาร รื่นชื่น)

..... อาจารย์ที่ปรึกษา
(อาจารย์ สลิลา ชีวกิดาการ)

..... ประธานสหกิจศึกษาสาขาวิชา
(อาจารย์ สลิลา ชีวกิดาการ)

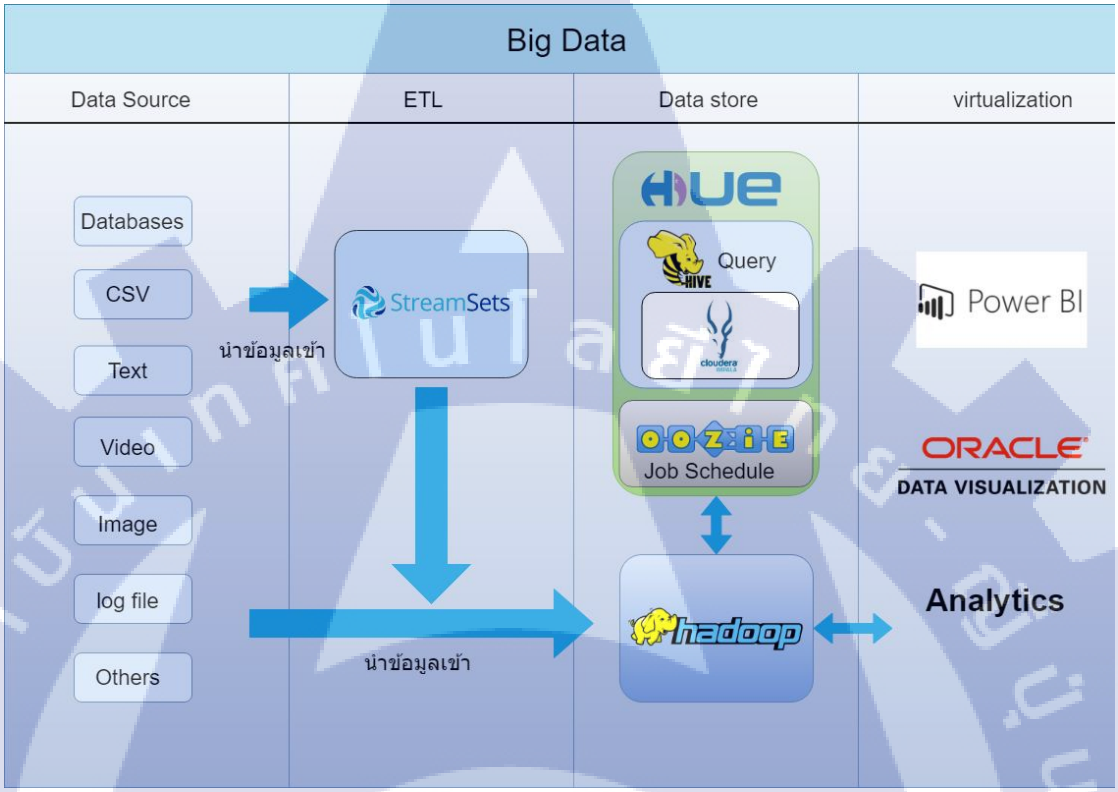
ชื่อโครงงาน	การดึงและรวบรวมข้อมูลขนาดใหญ่เพื่อนำมาประมวลผลให้เกิดประโยชน์และมีประสิทธิภาพสูงสุดด้วยเครื่องมือ Big Data Big Data gathering for the most effective analysis by using Big Data tools
ผู้เขียน	นาย ภูชิสส์ วัฒนกรวิโรจน์
คณะวิชา	เทคโนโลยีสารสนเทศ สาขาวิชา เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษา	อาจารย์ สลิลา ชีวกิดาการ
พนักงานที่ปรึกษา	นาย อภิสิทธิ์ แซ่ตั้ง
ชื่อบริษัท	บริษัท เอ-โฮสต์ จำกัด
ประเภทธุรกิจ/สินค้า	ให้บริการทางด้าน Oracle Product และ Hosting Service

บทสรุป

ในการสหกิจศึกษาได้รับมอบหมายในตำแหน่ง Assistant Technical Consultant ให้ช่วยในการสร้างระบบ Big Data ให้ บริษัท เอ - โฮสต์ จำกัด ที่มีวัตถุประสงค์ประสงค์เพื่อ ต้องการระบบที่สามารถจัดการDestributed file system ได้ และเครื่องมือต่างๆของ Big Data ที่สามารถใช้งานได้อย่างมีประสิทธิภาพสูงสุด

การติดตั้งระบบ Big Data นั้นได้ใช้เครื่องมือ Cloudera Manager ที่เป็นตัวจัดการให้ระบบ Big Data นั้นอยู่ในสภาพแวดล้อมเดียวกันเพื่อให้ Hadoop Destributed File System ซึ่งเป็นฐานข้อมูลของ Big Data นั้น สามารถนำไปใช้โดยเครื่องมือ Big Data ได้อย่างมีประโยชน์และประสิทธิภาพสูงสุด

Diagram Big Data



กิตติกรรมประกาศ

ในการที่ข้าพเจ้าได้มาสหกิจศึกษา ณ บริษัท เอ-โฮสต์ จำกัด ตั้งแต่วันที่ 4 มิถุนายน พ.ศ. 2561 ถึงวันที่ 28 กันยายน พ.ศ. 2561 ได้ทำให้ข้าพเจ้าได้เรียนรู้ประสบการณ์ต่าง ๆ ความรู้จากการทำงานจริง ซึ่งมีค่าอย่างเป็นอย่างมาก และส่งผลให้ข้าพเจ้าสามารถนำสิ่งต่าง ๆ ที่ได้จากการมาสหกิจ เหล่านั้นมาใช้พัฒนาทักษะของตนเอง สำหรับรายงานการปฏิบัติงานสหกิจศึกษาในครั้งนี้สามารถสำเร็จลุล่วงได้ด้วยดีจากความร่วมมือและการสนับสนุนจากหลายฝ่ายดังนี้

1. คุณบุญประสิทธิ์ ตั้งชัยสุข (Chief Executive Officer)
2. คุณสุชัย เย็นฤดี (Senior Vice President)
3. คุณเปรมสินี พิพัฒน์โกศล (HR Manager)
4. คุณพิชานน จะเรียมพันธ์ (Manager)
5. คุณนฤตยา สมศิริกุล (Programmer)
6. คุณอภิสิทธิ์ แซ่ตั้ง (Programmer)

และบุคลากรท่านอื่น ๆ ที่ไม่ได้กล่าวนามทุกท่านที่ได้ให้คำแนะนำช่วยเหลือในการจัดทำรายงานข้าพเจ้าขอขอบพระคุณ ผู้มีส่วนเกี่ยวข้องทุกท่านที่มีส่วนร่วมในการให้ข้อมูลเป็นที่ปรึกษาในการทำรายงานฉบับนี้จนเสร็จสมบูรณ์ ตลอดจนให้การดูแลและให้ความเข้าใจเกี่ยวกับชีวิตของการทำงานจริงข้าพเจ้าขอขอบพระคุณไว้ ณ ที่นี้

นาย ภูชิสส์ วัฒนกรวิโรจน์

ผู้จัดทำ

สารบัญ

สารบัญรูป

รูป **หน้า**

สารบัญตาราง

ตาราง **หน้า**

บทที่ 1
บทนำ

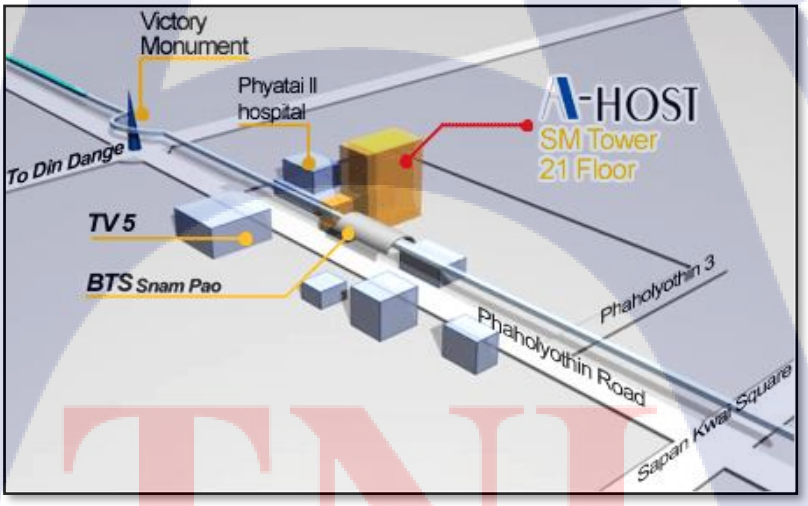
1.1 ชื่อและที่ตั้งของสถานประกอบการ

ชื่อ : บริษัท เอ-โฮสต์ จำกัด

ที่ตั้ง : 979/52-55 อาคาร เอส เอ็ม ทาวเวอร์ ชั้น 21 ถนนพหลโยธิน แขวงสามเสนใน เขตพญาไท กรุงเทพฯ 10400

ติดต่อ : 0-2298-0625-32 ต่อ 415 แฟกซ์ : 0-2298-0053

เว็บไซต์ : http://www.a-host.co.th



รูปที่ 1.1 แผนที่ตั้ง บริษัท เอ–โฮสต์ จำกัด

1.2 ลักษณะธุรกิจของสถานประกอบการ หรือการให้บริการหลักขององค์กร

บริษัทเอ-โฮสต์ก่อตั้งบริษัทเมื่อปี 2542 ซึ่งเป็นบริษัทหนึ่งในเครือของบริษัท เมโทร ซิสเต็มส์ คอร์เปอเรชั่น (มหาชน) จำกัด และเป็นผู้เชี่ยวชาญด้านบริการจัดวางระบบไอที และบริการเสริมต่าง ๆ สำหรับลูกค้าตั้งแต่ธุรกิจขนาดย่อมไปจนถึงขนาดกลาง

ธุรกิจหลักของบริษัท เอ-โฮสต์ คือ การให้บริการโฮสติ้ง และบริการระบบไอทีด้วยผลิตภัณฑ์ของออราเคิล (Oracle) และไอบีเอ็ม (IBM) ซึ่งเป็นซอฟท์แวร์สำหรับการวางแผนบริหารทรัพยากรขององค์กร (ERP) ระดับแนวหน้าของโลก

เอ-โฮสต์ถือกำเนิดขึ้นจากกลุ่มผู้เชี่ยวชาญด้านไอทีท่ามกลางภาวะเศรษฐกิจตกต่ำทั่วภูมิภาคเอเชียแต่ เอ-โฮสต์ ก็สามารถเติบโตได้อย่างรวดเร็ว และมั่นคงตั้งแต่แรกก่อตั้ง ด้วยจุดแข็งในฐานะผู้บุกเบิกธุรกิจโฮสติงเซอร์วิส พร้อมทั้งนำธุรกิจแนวใหม่อย่างการให้บริการแอพพลิเคชัน หรือ ASP (Application Services Providing) เข้ามาให้บริการแก่องค์กรธุรกิจเป็นรายแรกในเมืองไทย

ธุรกิจการให้บริการแอพพลิเคชัน ในรูปแบบ ASP ของเอ-โฮสต์ไม่เพียงแต่ให้บริการด้านแอพพลิเคชั่นด้านการดำเนินธุรกิจทางอิเล็กทรอนิกส์ระดับโลกของออราเคิลพร้อมโครงสร้างพื้นฐานทางเทคโนโลยีสารสนเทศเท่านั้น แต่ยังมีบริการที่ครอบคลุมตั้งแต่การให้คำปรึกษา การสนับสนุนและการให้บริการทั่วไปอย่างพรั่งพร้อมครบครัน รวมทั้งยังมีความยืดหยุ่นสูงมีการปรับเปลี่ยนบริการและทรัพยากรให้เหมาะสมกับความต้องการ และสภาพงานที่แตกต่างกันของลูกค้าในรายได้ระดับต่าง ๆ

ในการดำเนินธุรกิจของเอ-โฮสต์ตลอดระยะเวลา 10 ปี ไม่เพียงแต่ในฐานะผู้บุกเบิกธุรกิจโฮสติ้งและธุรกิจการให้บริการแอพพลิเคชั่นในรูปแบบ ASP เท่านั้น แต่เอ-โฮสต์ยังได้ทำการติดตั้งระบบไอที รวมทั้งผลิตภัณฑ์ของออราเคิลให้กับลูกค้าจนประสบความสำเร็จมาแล้วเป็นจำนวนมาก ซึ่งหลายรายเป็นหนึ่งในร้อยบริษัทชั้นนำของประเทศไทย แต่ที่สำคัญกว่านั้นก็คือการที่ เอ-โฮสต์ได้สานสัมพันธ์กับลูกค้า และพันธมิตรทางธุรกิจอย่างแนบแน่นจนกลายเป็นหุ้นส่วนทางกลยุทธ์และเป็นผู้สนับสนุนสำคัญที่มีส่วนช่วยผลักดันให้ธุรกิจของลูกค้าเติบโตสู่ความสำเร็จ

ปัจจุบันเอ-โฮสต์เป็นหนึ่งในบริษัทลูกของบริษัท ดราก้อนวัน จำกัด ซึ่งเป็นบริษัทมหาชนที่อยู่ในตลาดหลักทรัพย์แห่งประเทศไทย

เป็นเวลากว่า 10 ปีที่เอ-โฮสต์ และออราเคิลได้ดำเนินธุรกิจร่วมกันอย่างใกล้ชิด และถือเป็นพันธมิตรทางธุรกิจที่มีความแนบแน่นกันมานับตั้งแต่ก่อตั้งบริษัทปี 2542 จวบจนกระทั่งในปัจจุบัน

ในปี 2011-HOST ได้ก้าวไปข้างหน้าเพื่อความท้าทายทางธุรกิจใหม่ที่จะเป็น "พันธมิตรทางธุรกิจของไอบีเอ็มพรีเมียร์ (IBM Premier Business Partner)" เป็นตัวแทนจำหน่ายฮาร์ดแวร์ของไอบีเอ็มและผลิตภัณฑ์ซอฟต์แวร์ ที่จะสามารถเสริมสร้างผลิตภัณฑ์และผลงานบริการของเราเพื่อให้ลูกค้าได้รับเทคโนโลยีที่ดีที่สุดอยู่ตลอดเวลาและช่วยส่งเสริมการเจริญเติบโตของลูกค้าอย่างรวดเร็วและมั่นคง

ด้วยความมุ่งมั่นในการนำเสนอผลิตภัณฑ์ของออราเคิล และ ไอบีเอ็ม ผ่านการให้บริการแอพพลิเคชันในรูปแบบของ ASP ในฐานะที่เอ-โฮสต์เป็นผู้บุกเบิกธุรกิจดังกล่าว และเพิ่มศักยภาพในการดำเนินธุรกิจของลูกค้าได้อย่างเต็มประสิทธิภาพ และเหมาะสม ทำให้ได้รับรางวัลแห่งความสำเร็จและได้รับการยกย่องมาอย่างต่อเนื่อง

รูปที่ 1.2 A-HOST Proud Awards

1.3 รูปแบบการจัดองค์กรและการบริหารองค์กร

รูปที่ 1.3 คณะผู้บริหารบริษัท เอ-โฮสต์ จำกัด แต่ละแผนก

1.4 ตำแหน่งและหน้าที่งานที่นักศึกษาได้รับมอบหมาย

ตำแหน่งงานที่ได้รับมอบหมายในการปฏิบัติงานสหกิจ คือ Assistant Technical Consultant ขอบเขตงานที่ได้รับคือ ศึกษาเครื่องมือของ Bigdata เพื่อสามารถทำงานตาม Requirement โดยใช้เครื่องมือของ Bigdata ทำการโอนย้ายข้อมูลจาก Data Source สู่ Hadoop Distribute File System ผ่านกระบวนการ ETL โดยใช้เครื่องมือ StreamSets และ ทำการติดตั้งระบบ Cloudera เพื่อให้บริษัท A-HOST ใช้เครื่องมือของ Bigdata ได้อย่างเต็มประสิทธิภาพ

1.5 พนักงานที่ปรึกษา และ ตำแหน่งของพนักงานที่ปรึกษา

พนักงานที่ปรึกษา	:	นาย อภิสิทธิ์ แซ่ตั้ง
ตำแหน่ง	:	Programmer
E-mail	:	apisit@a-host.co.th

1.6 ระยะเวลาที่ปฏิบัติงาน

ปฏิบัติงานสหกิจเป็นเวลา 4 เดือน 1 สัปดาห์ตั้งแต่วันที่ 4 มิถุนายน 2561 ถึง วันที่ 28 กันยายน 2561 ก่อนสหกิจมีการอบรมเป็นเวลา 2 เดือน

1.7 ที่มาและความสำคัญของปัญหา

ในปัจจุบันหลายองค์กรในประเทศไทยและทั่วโลก กำลังกล่าวถึงข้อมูลขนาดใหญ่ Big Data เพราะองค์กรต่างๆทั้งในอดีตและปัจจุบัน ข้อมูลนั้นถือว่าเป็นสิ่งที่สำคัญในการบอกถึงสิ่งที่กำลังเกิดขึ้น บริษัท A-HOST จำกัด จึงเริ่มทำ Big Data โดยต้องการระบบ Big Data และเครื่องมือต่าง ๆ ของ Big Data และวิธีการใช้งาน

1.8 วัตถุประสงค์หรือจุดมุ่งหมายของโครงงาน

1. เพื่อศึกษาหาความรู้ความเข้าใจในการปฏิบัติงานด้าน Big Data
2. เพื่อเข้าใจกระบวนการการทำงานของเครื่องมือ Big Data
3. เพื่อนำ Big Data ต่อยอดใช้ในธุรกิจได้

1.9 ผลที่คาดว่าจะได้รับจากการปฏิบัติงานหรือโครงงานที่ได้รับมอบหมาย

1. นักศึกษานำความรู้ที่ได้จากการปฏิบัติงานสหกิจไปใช้ประกอบอาชีพในอนาคต
2. นักศึกษาสามารถทำงานร่วมกันผู้อื่นได้เป็นอย่างดี
3. นักศึกษามีความรู้และทักษะเฉพาะทางในสายงานนี้มากยิ่งขึ้น
4. นักศึกษามีความรับผิดชอบในหน้าที่การงานที่ได้รับมอบหมาย
5. ระบบสามารถใช้งานได้จริง
6. ระบบสามารถโอนย้ายข้อมูลตามความต้องการของลูกค้าได้
7. ระบบมีการโอนย้ายข้อมูลได้อย่างครบถ้วน

1.10นิยามศัพท์เฉพาะ

1. BI (Business Intelligence) หมายถึง ชุดของแนวคิดและกระบวนทัศน์ที่จะพัฒนากระบวนการ ตัดสินใจของธุรกิจโดยอาศัยข้อมูลที่เป็นข้อเท็จจริงจากฐานข้อมูล

2. Bigdata หมายถึง ปริมาณข้อมูลจำนวนมหาศาลที่มีอยู่ในบริษัทของคุณทุกรูปแบบ ไม่ว่าแหล่งที่มาจะมาจากภายในบริษัทหรือภายนอกก็ตาม ทั้งนี้แบ่งออกเป็นข้อมูลที่มีโครงสร้างชัดเจน (Structured Data) และข้อมูลที่มีโครงสร้างไม่ชัดเจน (Unstructured Data)

3. StreamSets คือ เครื่องมือสำหรับการดึงข้อมูล แปลงข้อมูลและโหลดข้อมูล

4. HDFS หรือ Hadoop Distribution File System คือ เป็นการเก็บข้อมูลแบบกระจาย โดยการหั่น ไฟล์เป็นblock และกระจายไปตามCluster ต่าง ๆของเรา

5. Cloudera คือ software open source ที่เป็นตัวจัดการTools ต่างๆของbigdata ให้อยู่ในสภาพแวดล้อมเดียวกัน และเป็นตัวลง Tools ต่างๆของBigdata ให้User ง่ายต่อการจัดการมากขึ้น

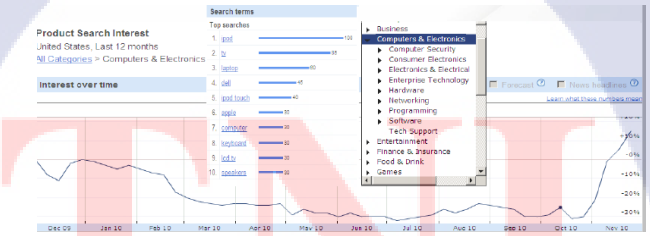
บทที่ 2
ทฤษฎีและเทคโนโลยีที่ใช้ในการปฏิบัติงาน

ในการปฏิบัติงานสหกิจศึกษาครั้งนี้ เป็นการนำความรู้ทางด้านทฤษฎีและเทคโนโลยีมาใช้ในการปฏิบัติงานทุกส่วนตลอดการปฏิบัติงานสหกิจศึกษา ซึ่งเป็นการนำความรู้ทั้งที่เคยเรียนมาประยุกต์ใช้และเป็นการศึกษาเรียนรู้สิ่งใหม่ ๆ ที่ได้จากการปฏิบัติงาน

2.1 ระบบปัญญาธุรกิจ (Business Intelligence: BI)

BI คือ การนำเสนอข้อมูลเพื่อช่วยในการวางแผนการตัดสินใจหรือตอบคำถามเชิงธุรกิจให้กับผู้บริหาร ทำให้องค์กรสามารถคาดการณ์หรือพยากรณ์ความต้องการของผู้บริโภคได้อย่างถูกต้องแม่นยำ ส่งผลให้ประสิทธิภาพการทำงานขององค์กรสูงขึ้น

Business Intelligence หรือที่เรียกว่า “BI” นั้นได้ถูกคิดค้นโดย Howard Dresner ในช่วงปี 1990 Howard Dresner ให้ความหมาย คำว่า “ระบบปัญญาธุรกิจ” (Business Intelligence) หรือ BI หมายถึง “ชุดของแนวคิดและกระบวนทัศน์ที่จะพัฒนากระบวนการตัดสินใจของธุรกิจโดยอาศัยข้อมูลที่เป็นข้อเท็จจริงจากฐานข้อมูล”



รูปที่ 2.1 บทบาทของ BI

2.1.1 ลักษณะสำคัญของระบบ BI คือ

ระบบงานของ BI

จุดเด่นของ BI คือ ใช้งานง่ายผู้ใช้ไม่ต้องมีความรู้ด้านฐานข้อมูล ก็สามารถใช้งานได้ เพียงเลือกข้อมูลที่ต้องการก็สามารถได้ผลลัพธ์ตามต้องการ ข้อมูลมีความถูกต้องแม่นยำทำให้สามารถใช้ข้อมูลเพื่อช่วยในการตัดสินใจได้รวดเร็วกว่าคู่แข่ง ทั้งในเชิงกว้าง และเชิงลึก สามารถดึงข้อมูลจากฐานข้อมูลที่หลากหลายมาทำการวิเคราะห์โดยไม่มีการเขียนโปรแกรม

รูปที่ 2.2 ระบบงานของ BI

2.1.2 กระบวนการทำงานของ BI

ก่อนที่จะเกิดเป็น BI ได้ต้องมีกระบวนการทำงานต่าง ๆ ที่ทำงานร่วมกันเพื่อให้ข้อมูลออกมาตามที่ต้องการและวางแผนไว้ล่วงหน้าได้โดนต้องมีกระบวนการต่าง ๆ ดังต่อไปนี้

1) การกำหนดแหล่งข้อมูล (Data Sources) แบ่งออกเป็น 2 ประเภท คือ

- แหล่งข้อมูลภายใน (Internal Data Sources) ได้แก่ ข้อมูลการดำเนินงาน (Operation Transaction) ข้อมูลอดีต (Legacy Data) เป็นต้น

- แหล่งข้อมูลภายนอก (External Data Sources) ได้แก่ ข้อมูลสถิติจากสถาบันต่าง ๆ ข้อมูลของโครงการสารสนเทศอื่น ๆ บทวิเคราะห์และบทความวิชาการต่าง ๆ

2) การออกแบบคลังข้อมูล (Data Warehouse Design)

- เป็นที่จัดเก็บข้อมูลที่นำมาจากแหล่งข้อมูลภายในองค์กรจากฐานข้อมูลการใช้งานประจำวันหรือฐานข้อมูลปฏิบัติการ (Operational Database) หรือมาจากฐานข้อมูลภายนอกองค์กร (External Database)

- ข้อมูลในคลังข้อมูลจะถูกนำมาใช้เพื่อสนับสนุนการตัดสินใจบริหารงานของผู้บริหาร

- เช่น ระบบสนับสนุนการตัดสินใจ (Decision Support System) หรือ ระบบการจัดการสารสนเทศ (Management Information System)

รูปที่ 2.3 Data Warehouse

รูปแบบการนำข้อมูลมาประมวลผลแบ่งออกเป็น 2 ประเภทใหญ่ ๆ ดังนี้

OLTP (Online Transaction Processing) คือ การประมวลผลข้อมูลตามลักษณะการปฏิบัติงานประจำวันของหน่วยงานนั้นจากฐานข้อมูลเช่น ERP (Enterprise resource planning), POS (point of sale)

OLAP (Online Analysis Processing) เป็นเทคโนโลยี ที่ใช้ดึงข้อมูลจาก Data Warehouse นำไปวิเคราะห์เพื่อใช้สนับสนุนการตัดสินใจของผู้บริหารจากคลังข้อมูลรูปแบบในการวิเคราะห์มีดังนี้

- BI (Business Intelligence) ซอฟต์แวร์ที่นำข้อมูลที่มีอยู่เพื่อจัดทำรายงานในรูปแบบต่าง ๆ ที่เหมาะสมกับมุมมองในการวิเคราะห์และตรงตามความต้องการของผู้ใช้งานตามแต่ละแผนก
- Data mining ระบบช่วยดูแนวโน้มในอนาคตความสัมพันธ์ของข้อมูล
- DSS (Decision Support System) ระบบสนับสนุนผู้บริหารเพื่อช่วยผู้บริหารในการตัดสินใจเชิงกลยุทธ์

รูปที่ 2.4 OLAP (online analysis processing)

OLAP แบบเป็น 3 ประเภท

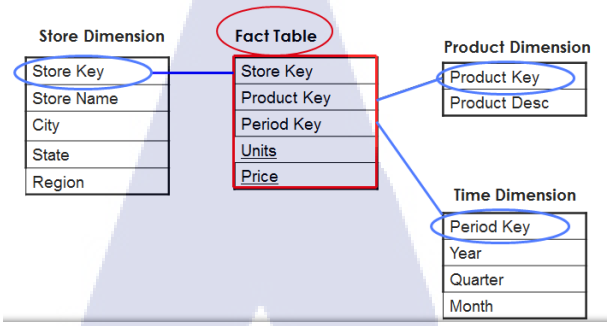
- ROLAP (Relational OLAP) คือ OLAP หรือ Cube ที่ไม่จำเป็นต้องมีการประมวลผล OLAP ไว้ก่อน แต่จะเก็บข้อมูลในรูปแบบของ Relational Database เมื่อมีการเรียกใช้ข้อมูลจาก ROLAP ระบบจะไปดำเนินการสร้าง Query เพื่อดึงข้อมูลออกมาจาก Fact Table วิธีการนี้จะช้ากว่าแบบ MOLAP แต่ข้อมูลที่ได้ทันสมัยเสมอ

- MOLAP (Multidimensional OLAP) คือ OLAP หรือ Cube ที่จะต้องมีการประมวลผล Fact Table เพื่อใส่ค่าในช่องต่าง ๆ ของ Cubeไว้ ก่อนที่จะมีการใช้งานค่าในแต่ละช่องของ MOLAP จะคงที่ไม่เปลี่ยนแปลง จนกว่าจะมีการประมวลผลใหม่อีกครั้ง แต่การเรียกใช้งานจาก MOLAP จะรวดเร็วมาก

- HOLAP (Hybrid OLAP) คือ OLAP หรือ Cube ที่มีการแบ่งพื้นที่ออกเป็นส่วน ๆ โดยแต่ละส่วนอาจใช้วิธีการจัดเก็บข้อมูลแบบ MOLAP และบางส่วนก็จัดเก็บแบบ ROLAP

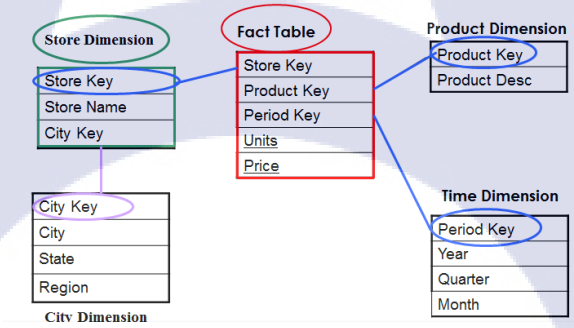
Design Data Warehouse ซึ่งการออกแบบคลังข้อมูลมีอยู่ 2 แบบ

การออกแบบคลังข้อมูลแบบ Star Schema หรือ Multidimensional Schema เป็น Dimensional data ที่ประกอบไปด้วยตารางสองชนิดด้วยกัน คือ Fact Table และ Dimension Table โครงสร้าง Star Schema จะประกอบไปด้วย Fact Table อยู่ตรงกลาง และล้อมรอบไปด้วย Dimension Table เพื่อกำหนดมุมมองที่จะมีต่อข้อมูลใน Fact Table นั้นโดย Fact Table จะเป็นศูนย์รวมข้อมูลเพียงTable เดียวและจำนวนมุมมองจะเท่ากับจำนวนของ Dimension ที่รายล้อมอยู่ ซึ่งลักษณะแบบนี้ จะช่วยเพิ่มความสามารถในการ Query ข้อมูลได้ง่ายและรวดเร็ว



รูปที่ 2.5 คลังข้อมูลแบบ Star Schema หรือ Multidimensional Schema

คลังข้อมูลแบบ Relational Schema หรือ Snowflake Schema หมายถึง Dimensional Data Model ที่มี Fact Table ขนาดใหญ่เพียงหนึ่งเดียวอยู่ตรงกลาง และมี Dimensional Table จำนวนหนึ่งอยู่รายรอบ โดยที่ Dimension Table ที่รายรอบอยู่นั้น สามารถที่จะเชื่อมต่อไปยัง Dimension Table อื่น ๆ ได้อีกดังนั้นโครงสร้างแบบนี้จะซับซ้อนมากขึ้น รวมทั้งมีผลให้การใช้ Query ยากขึ้นอีกด้วย



รูปที่ 2.6 Snowflake Schema

องค์ประกอบสำคัญในตารางมี 2 ประเภท คือ

- Fact Table เป็นตารางหลัก ซึ่งมีลักษณะคล้ายกับตารางประเภท Transaction ของ OLTP โดยภายในจะประกอบด้วยคอลัมน์ที่สำคัญ 2 ประเภทคือ

- Key เป็นคอลัมน์ที่ใช้เชื่อมโยงไปยัง Dimension Table ต่าง ๆ ดังนั้นจำนวนคอลัมน์ของ Fact Table Key จะเพิ่มขึ้นตามจำนวนของ Dimension Table และคอลัมน์ทั้งหมดนี้สามารถนำไปใช้สร้างให้เป็น Primary Key ได้อีกด้วย

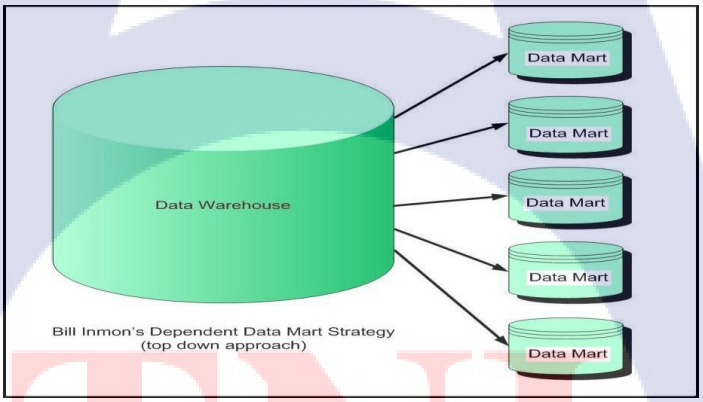
- Measure เป็นข้อมูลตัวเลข ทำหน้าที่เก็บจำนวน หรือปริมาณที่เกิดขึ้นของแต่ละ Transaction นอกจากนี้ยังเก็บผลลัพธ์ที่ได้จากการคำนวณ

Dimension Tables เป็นตารางข้อมูลที่เก็บการอธิบาย Entity ต่าง ๆ โดยประกอบด้วยคอลัมน์ที่เป็น Key เพื่อเชื่อมโยงไป Fact Table Key และคอลัมน์ที่ให้ความหมายเพิ่มเติมแก่ Entity สามารถนำไปสร้างเป็น Dimension ของ OLAP Cube ตามลักษณะต่าง ๆ ดังนี้

- Standard Dimension มาจาก Dimension Table ปกติ ซึ่งแต่ละคอลัมน์อธิบายข้อมูล Entity นั้น ๆ เพียงอย่างเดียว

- Parent-Child Dimension มีลักษณะคล้ายกับ Standard Dimension แต่ภายในจะมีความสัมพันธ์ระหว่างภายในกันเอง

Data Mart คือ คลังข้อมูลขนาดเล็กมีการเก็บข้อมูลที่มีลักษณะเฉพาะเจาะจง เช่น เก็บข้อมูลส่วนของการเงิน ส่วนของสินค้าลงคลัง ส่วนของการขาย เป็นต้น ซึ่งทำให้การจัดการข้อมูลการนำเอาข้อมูลไปสร้างความสัมพันธ์และวิเคราะห์ต่อ ก็ง่ายขึ้น เมื่อมีข้อมูลใน Data Warehouse แล้วสามารถสร้าง Data Mart ขึ้นมาได้ ซึ่ง Data Mart นั้นถูกออกแบบมาเพื่อช่วยเหลือผู้ใช้ในการตอบคำถามทางธุรกิจตามที่ผู้ใช้ต้องการ ขั้นตอนการย้ายข้อมูลจาก Data Warehouse เข้าสู่ Data Mart เรียกว่า Data Deliver



รูปที่ 2.7 Data Mart

3) การคัดเลือก (ETL Extract-Transform-Load)

ETL ย่อมาจาก Extract, Transform, and Load คือการดึงข้อมูลจาก Data Source ต่าง ๆ เข้าสู่ Data Warehouse แบ่งเป็น 5 ขั้นตอนหลัก ดังนี้

- Extract คือ การดึงข้อมูลจากแหล่งข้อมูลที่แตกต่างกัน

- Transform คือ การนำข้อมูลมาจัดรูปแบบให้ถูกต้องสอดคล้องกัน

- Data Mapping การทำให้ข้อมูลให้อยู่ในรูปแบบเดียวกัน

- Data Cleansing การตรวจสอบและแก้ไขข้อมูลให้ถูกต้อง

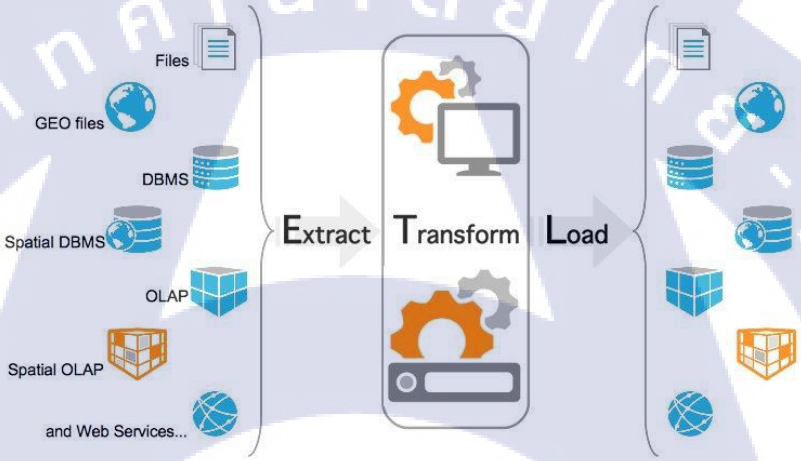
- Load คือ การนำข้อมูลที่ผ่านการ Transform แล้ว เข้าสู่ Data Warehouse

โดยทั่วไปแล้ว ETL Tools สามารถทำได้ดังนี้

- Data Cleansing-ตรวจสอบความถูกต้องของข้อมูล รวมทั้งกำจัดข้อมูลที่ผิดพลาด

- Data Transformation–การแปลงข้อมูล หรือจัดรูปแบบข้อมูลเพื่อให้สามารถนำไปวิเคราะห์ได้ง่ายขึ้น

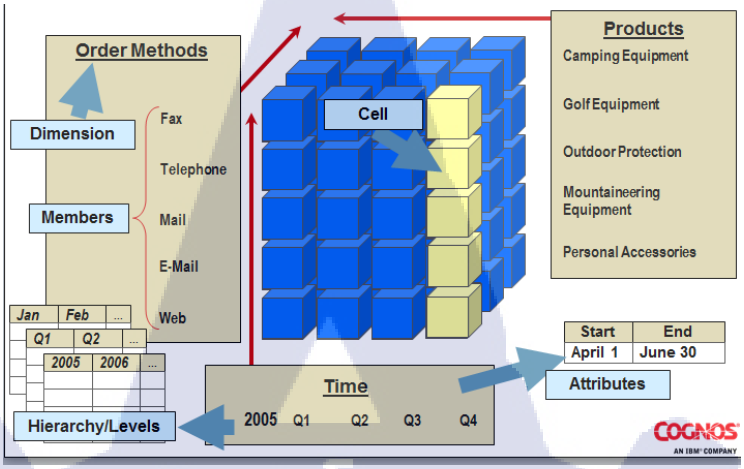
- Data Loading and Refreshing-กำหนด schedule ได้ว่าจะให้โหลดมาทุก ๆ กี่วัน หรือทุก ๆ เท่าไหร่ รวมทั้งยังสามารถกำหนด storage ปลายทางได้อีกด้วย



รูปที่ 2.8 ETL Extract-Transform-Load

4) การทำข้อมูลให้อยู่ในรูปแบบ Multidimensional Model หรือ Cube

รูปแบบการทำให้ข้อมูลเกิดมิติขึ้นในหลาย ๆ ด้านก่อนจะนำไปสร้างเป็นรายงานในรูปแบบต่าง ๆ โดยอาศัยเครื่องมือที่ช่วยในการ Query ข้อมูลที่ต้องการ เช่น Oracle Essbase, Cognos Transformer



รูปที่ 2.9 Cube

5) การออกแบบและสร้างรายงาน (Interactive Report)

รายงานที่นำเสนอมักจะเป็นผลการดำเนินงานตามตัวบ่งชี้การดำเนินงานต่าง ๆ ของหน่วยงานหรือการติดตามค่าเป้าหมายของการดำเนินงาน ที่สำคัญการนำเสนอรายงานมักจะอยู่ในรูปแบบของกราฟเพื่อทำให้เกิดความเข้าใจได้ง่ายผ่าน Dashboard ที่ผู้ใช้สามารถเข้าถึงผ่านหน้าเว็บไซต์ที่จัดทำไว้ เช่น OBI Report, Cognos Report เป็นต้น

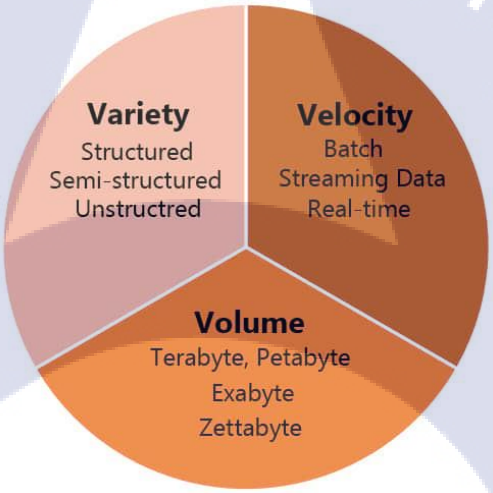
รูปที่ 2.10 กระบวนการทำงานของ BI

2.1.3 ประโยชน์ของ BI

- ช่วยเพิ่มศักยภาพในการตัดสินใจให้ถูกต้องและรวดเร็วจากข้อมูลที่มีอยู่ โดยเห็นภาพพจน์ของข้อมูลที่มีก่อนการตัดสินใจ
- เพิ่มประสิทธิภาพในการแลกเปลี่ยนข้อมูลภายในองค์กร โดยสามารถแลกเปลี่ยนข้อมูลภายในผ่านเครือข่ายได้ในแบบอัตโนมัติ
- ลดต้นทุนทั้งด้านเงินและเวลาในการเข้าถึงข้อมูลองค์กร
- ช่วยให้ผู้ใช้สามารถตอบคำถามทางธุรกิจที่สำคัญ และช่วยให้สามารถ รวบรวมและปรับข้อมูลตามต้องการเพื่อสร้างแนวคิดที่แตกต่าง
- สำรวจข้อมูลทุกชนิดจากทุกแง่มุม
- วิเคราะห์ข้อเท็จจริงและคาดการณ์นัยแฝงเชิงยุทธวิธีและเชิงกลยุทธ์

2.2 ข้อมูลขนาดใหญ่ (Bigdata)

Bigdata คือ ปริมาณข้อมูลที่มีขนาดใหญ่มหาศาลเกินกว่าขีดความสามารถในการประมวลผลของระบบฐานข้อมูลธรรมดาที่จะรองรับได้ปริมาณข้อมูลที่มีขนาดใหญ่มาก ๆ จะมีอัตราการเพิ่มข้อมูลได้อย่างรวดเร็วมากและจะมีรูปแบบที่มีโครงสร้างหรือไม่มีโครงสร้าง ซึ่งไม่สามารถอยู่ในระบบฐานข้อมูลที่จะจัดเก็บข้อมูลที่ได้ โดย Bigdata นั้นมีลักษณะพื้นฐานอยู่ 3 ลักษณะหลัก ๆ อธิบายโดยใช้ 3V Model ได้แก่ ปริมาณ (Volume) ความเร็ว (Velocity) และความหลากหลาย (Variety)



รูปที่ 2.11 3V Model of Big data

ปริมาณ (Volume) คือปริมาณของข้อมูลที่มีจำนวนมหาศาล ซึ่งในอดีตการเก็บข้อมูลที่มีขนาดใหญ่นั้นมักจะเป็นปัญหา แต่ในปัจจุบันเทคโนโลยี Big Data ทำให้การจัดการทำข้อมูลที่มีขนาดใหญ่นี้เป็นเรื่องที่ง่ายขึ้นกว่าเดิมมาก

ความเร็ว (velocity) คือการเข้าถึงข้อมูลและจัดการข้อมูลนั้นจะต้องทำได้แบบทันทีทันใด ข้อมูลที่เกิดขึ้นมาควรถูกเก็บและวิเคราะห์ให้ตรงเวลา

ความหลากหลาย (Variety) คือรูปแบบของข้อมูลที่แตกต่างกันออกไป ไม่ว่าจะเป็นตัวอักษร วิดีโอ รูปภาพ หรือก็คือข้อมูลแบบมีโครงสร้าง และอื่นๆอีกมากมาย

2.3 เทคโนโลยีที่ใช้ในการปฏิบัติงาน

2.3.1 โปรแกรม Oracle Virtual Box

รูปที่ 2.12 Logo Oracle Virtual Box

Virtual Box (ชื่อเต็มคือ Oracle VM Virtual Box) เป็นโปรแกรมฟรีแวร์สำหรับจำลองระบบคอมพิวเตอร์ เป็นซอฟต์แวร์สำหรับใช้ทำการจำลองระบบคอมพิวเตอร์ (Virtualization) บนระบบ x86 และ AMD64/Intel64 ลักษณะเดียวกับโปรแกรม VMware Workstation (เป็นโปรแกรมเชิงพาณิชย์ต้องซื้อจึงจะใช้งานได้เต็มฟังก์ชัน) และ VMware Player 3.0 (สามารถใช้งานได้ฟรี) ของVMware หรือโปรแกรม Virtual PC ของ Microsoft ซึ่งสามารถใช้งานได้ฟรี และ Windows Virtual PC ของ Microsoft ซึ่งก็จะสามารถใช้งานได้ฟรีแต่จะมีเฉพาะใน Windows 7 รุ่น Professional, Enterprise และ Ultimate

Virtual Box เป็นซอฟต์แวร์แบบ Open Source พัฒนาโดย Oracle (ก่อนหน้านี้เป็น Sun Microsystems) ซึ่งปัจจุบันถูกซื้อกิจการโดย Oracle สามารถใช้งานได้ฟรีโดยไม่มีค่าใช้จ่ายภายใต้ไลเซนส์แบบ GNU General Public License (GPL) เป็นซอฟต์แวร์ที่มีประสิทธิภาพสูงรองรับการใช้งานได้ทั้งในเอนเทอร์ไพรส์ (Enterprise) และการใช้งานภายในบ้าน และยังมีฟีเจอร์ให้ใช้งานหลากหลายและที่สำคัญเป็นโซลูชั่นระดับมืออาชีพที่ใช้งานได้ฟรี

Virtual Box คือโปรแกรมที่ใช้ในการจำลองเครื่องคอมพิวเตอร์ขึ้นมาอีกเครื่องหนึ่งโดยการแบ่งทรัพยากรจากระบบหลักไปใช้เช่น CPU, RAM, VGA, และ HDD โดยจุดมุ่งหมายหลักของโปรแกรมนี้คือการติดตั้ง ระบบปฏิบัติการขึ้นมาอีกตัวหนึ่งเพื่อใช้งานที่แตกต่างกันไป

2.3.2 โปรแกรม Cloudera

รูปที่ 2.13 Logo Cloudera

Cloudera เป็นโปรแกรม open source สำหรับทำงานกับ Hadoop เท่านั้น ที่จะทำให้ระบบของHadoop ทำงานอย่างมีประสิทธิภาพที่ดีที่สุด ทำให้การขับเคลื่อนข้อมูลขนาดใหญ่ หรือ Bigdata ให้เกิดผลลัพท์ที่ต้องการได้อย่างรวดเร็ว คุณลักษณะที่สำคัญ ได้แก่

- การประมวลผลข้อมูลภายในหน่วยความจำ : ด้วยTools Apache Spark

- การวิเคราะห์ที่รวดเร็วด้วย SQL: ค่าแฝงต่ำสุด เห็นพ้องที่สุดสำหรับโซลูชันการวิเคราะห์ข้อมูลเพื่อเพิ่มประสิทธิผลทางธุรกิจด้วย Apache Impala

- การค้นหาแบบดั้งเดิม: การเข้าถึงข้อมูลของผู้ใช้ได้อย่างสมบูรณ์ซึ่งติดตั้งมาพร้อมกับแพลตฟอร์ม Apache Solr

- อุปกรณ์เก็บข้อมูลเชิงวิเคราะห์ที่ปรับปรุงให้ทันสมัยอยู่เสมอ: อุปกรณ์จัดเก็บข้อมูล Hadoop ที่สามารถวิเคราะห์ข้อมูลที่เปลี่ยนแปลงได้อย่างรวดเร็วเมื่อใช้งานร่วมกับ Apache Kudu เท่านั้น

- การปรับแต่งข้อมูลเชิงรุกให้เกิดประสิทธิผลสูง: Cloudera Navigator Optimizer (limited beta) ช่วยในการปรับแต่งข้อมูลและปริมาณงานเพื่อประสิทธิภาพการดำเนินงานสูงสุดในการทำงานร่วมกับ Hadoop

2.3.3 Apache Hadoop

รูปที่ 2.14 Logo Hadoop

Hadoop เป็นซอฟท์แวร์ประเภท open source ที่จัดทำขึ้นเพื่อเป็นแพลตฟอร์มในการจัดเก็บข้อมูล ซึ่งมีกรอบการทำงานเพื่อใช้ในการจัดเก็บข้อมูลและประมวลผลข้อมูลที่มีขนาดใหญ่ ที่เราเรียกกันว่า Big Data ซึ่ง Hadoop ก็สามารถปรับขยาย ยืดหยุ่น เพื่อรองรับข้อมูลที่มีจำนวนมากมายมหาศาลได้ ทั้งนี้ก็เพราะมันมีการกระบวนการประมวลผลที่แข็งแกร่งมากซึ่งเป็นผลมาจากการประมวลผลข้อมูลแบบกระจายผ่านเครื่องคอมพิวเตอร์ที่ถูกจัดอยู่ในรูปแบบ Cluster อันนำไปสู่ความสามารถในการรองรับข้อมูลที่ไม่จำกัดแถมยังมีความน่าเชื่อถือสูงอีกด้วย

ประวัติความเป็นมาของ Hadoop ต้องย้อนกลับไปในปี 2006 หลังจากที่ World Wide Web เติบโตจนถึงจุดที่การใช้งานอินเตอร์เน็ตมีการขยายวงกว้างออกไปเรื่อยๆ ผู้ค้นค้นหาข้อมูลต่างๆ พอๆกับที่มีการป้อนคอนเท้นท์และข้อมูลเข้าไป 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

องค์ประกอบหลักของ Hadoop คือ Hadoop Distributed File System (HDFS) ตัวเก็บข้อมูลของ Bigdata และ MapReduce (Yarn) สำหรับประมวลผล

2.3.4 โปรแกรม Hue (Hadoop User Experience)

รูปที่ 2.15 หน้าโปรแกรม HUE

Hue หรือ Hadoop User Experience

เป็นหน้าจอ Web UI ของค่าย Cloudera ทำให้สามารถจัดการ Application ตัวอื่นๆ เช่น Hive,impala,pig,sqoop,Hbase,HDFS เป็นต้น ผ่านทาง Web Browser ได้อย่างสะดวกและง่ายกว่าใช้คำสั่ง Command Line

2.3.5 Apache Hive

รูปที่ 2.16 Logo HIVE

Hive

เป็นเครื่องสำหรับผู้ต้องการสืบค้นข้อมูล (Query) ข้อมูลที่เก็บใน HDFS โดยใช้ภาษาที่เรียกว่า Hive QI ซึ่งมีลักษณะคล้ายภาษา SQL แทนที่การเขียนโปรแกรม Map/Reduce เนื่องจาก Hive จะทำการแปลง Hive QL เป็น Map/Reduce แล้วทำการ Execute เป็นแบบ Batch

2.3.6 โปรแกรม Impala

รูปที่ 2.17 Logo Impala

Impala

เป็นเครื่องมือที่ทางค่าย Cloudera ทำการ Build เข้ามาในตัว Cloudera Hadoop นั่นเอง โดยมีการทำงานคล้ายๆกับ Hive แต่ที่แตกต่างกันคือ Impala จะทำงานกับข้อมูลที่อยู่บนMemory และติดต่อข้อมูล HDFS โดยตรงโดยที่ไม่ต้องผ่าน MapReduce ซึ่งจะทำให้Impala เร็วกว่าHive อย่างแน่นอน

แต่ Hive มีคุณสมบัติ Fault Tolerance เช่น ระบบกำลังทำงานแล้วระบบล่ม ถ้ากู้กลับมาแล้ว Hive ทำงานต่อได้ แต่ Impala ต้องมาสั่งทำงานใหม่อีกรอบครับ

2.3.7 โปรแกรม Ozzie

รูปที่ 2.18 Logo OOZIE

Ozzie

เป็นเครื่องมือในการทำ Workflow จะช่วยให้เราเอาคำสั่งประมวลผลต่างๆของระบบ Hadoop เช่น Map/Reduce, Hive หรือ Pig มาเชื่อมต่อกันในรูปของ Workflow ได้

2.3.8 โปรแกรม Kafka

รูปที่ 2.19 Logo Kafka

Kafka

เป็น Streaming Platform ตัวหนึ่งทำหน้าที่เป็น Broker รับ Message จากที่นึงไปยังอีกที่หนึ่ง โดย Message จะเป็น record ผ่าน TCP protocal ไปมา เพื่อช่วยในการ scalability & decoupling ของระบบ ให้มัน asynchronus มากที่สุด

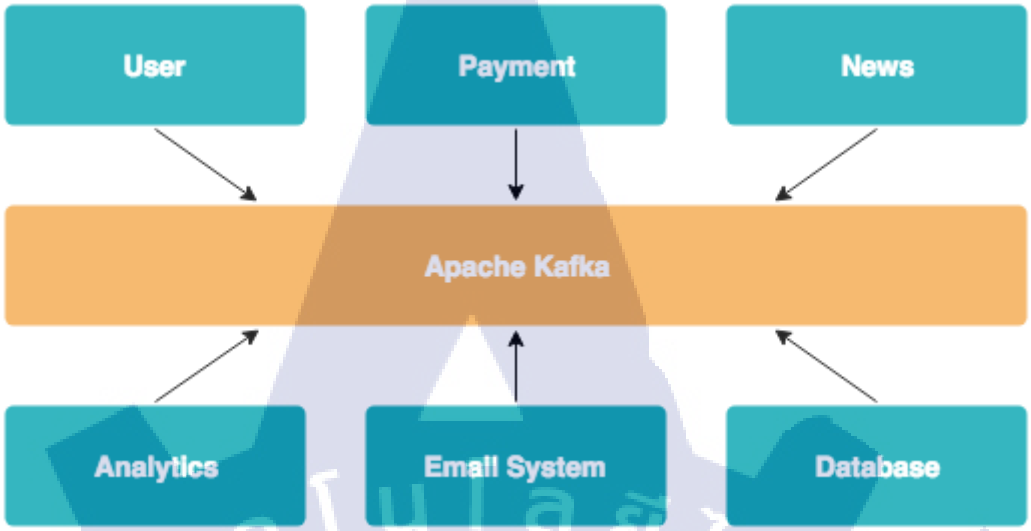
โดยเริ่มแรก Kafka ถูกสร้างขึ้นโดย LinkedIn เป็น open source project ด้วยภาษา Java และ Scalaในช่วงต้นปี 2011 ที่ในสมัยก่อนเค้าต้องจัดการกับปัญหาเรื่องของ

1. ผู้ใช้จำนวน 300 million users events ทุกวันๆ ทำให้ในบางทีมันเกิดปัญหาเรื่องของ data lost(รวมไปถึงการทำ internal services messaging ด้วย นึกถึงสภาพบริษัทเรามี cluster A, cluster B, cluster C เยอะแยะไปหมด จะทำการติดต่ออย่างไรให้มีประสิทธิภาพ)
2. ปัญหาเรื่องของ การย่อยข้อมูลขนาดใหญ่ ขนาดนั้นเวลาเอามาใช้งานอีกด้วย

จึงเกิดโปรเจคนี้ขึ้นมาและถูกเผยแพร่ต่อ ผ่านทาง Apache Incubator ตั้งแต่ปี 2012 จากนั้นจึงได้แยกบริษัทออกมาจาก LinkedIn ก่อตั้งเป็น บริษัท Confluent เพื่อพัฒนา Kafka โดยเฉพาะ โดยชื่อ Kafka มาจากนักเขียนนาม Franz Kafka โดยเลือกชื่อ Kafka เพราะมันถูก optimize สำหรับการเขียน เหมือนกับงานของ Franz Kafka

Kafka ดีอย่างไร

ยกตัวอย่าง ถ้ามีระบบสองตัวให้ติดต่อกันระหว่างสองตัวนี้ ก็จะติดต่อกันแบบ peer to peer แต่ถ้าหากมีมากขึ้นก็ต้องทำการเชื่อมต่อในแต่ละระบบมากขึ้น



รูปที่ 2.20 ตัวอย่างserverส่งข้อมูลเข้า Kafka

ดังนั้น Apache Kafka จึงถูกใช้เป็นตัวกลางในการเชื่อมต่อ เพื่อช่วยแยกการสื่อสารระหว่างระบบแต่ละตัว ไม่ว่าระบบอะไรต้องการข้อมูลจากไหน ก็มาเรียกใช้ใน Apache Kafka ตัวนี้ตัวเดียว

คุณสมบัติของ Apache Kafka

- มีการกระจาย (distributed) การเก็บข้อมูลใน clusters
- มีความยืดหยุ่น (resilient architecture) เช่น มีการทำสำเนาข้อมูลซ้ำ (replication)
- มีการทนต่อความเสียหาย (fault tolerent)
- มีความสามารถในการขยายเชิงขนาน หรือ เพิ่มเครื่อง (node) ใน cluster ได้ (horizontal scalability)
- มีประสิทธิภาพด้านความเร็ว (latency น้อยกว่า 10ms)
- มีระบบที่ใหญ่ๆ ที่ใช้ Apache Kafka อยู่มาก เช่น Linkedin, Netflix, AirBnB, Yahoo, Wallmart หรือ LINE

สิ่งที่ Apache Kafka สามารถทำได้

- ระบบส่งต่อข้อความ (messaging System)
- เครื่องมือบันทึกกิจกรรม (activity Tracking)
- รวบรวมเก็บ Log (log aggregation)
- การประมวลผลแบบต่อเนื่องของข้อมูล (stream processing)

บทที่ 3
แผนงานการปฏิบัติงานและขั้นตอนการดำเนินงาน

3.1 แผนงานการฝึกงาน

ตารางที่ 3.1 ตารางแผนการฝึกงาน

หัวข้องาน	เดือนที่ 1				เดือนที่ 2				เดือนที่ 3				เดือนที่ 4			
ศึกษา Oracle Business Intelligence Enterprise																
ศึกษา Concept Big Data																
Installation Program Linux สำหรับทดสอบ Big Data																
Install LibreOffice on Linux																
แปลงข้อมูลผ่านเครื่องมือ LibreOffice โดยใช้ Script แปลง																
ศึกษา Python และ Modify Script																
ศึกษาเครื่องมือ ETL StreamSets																
ศึกษาเครื่องมือทางด้าน Big Data ได้แก่ Hue,hive และImpala																
ศึกษาเครื่องมือต่อยอดเพิ่มเติมจากโครงการ Big Data ได้แก่ Ozzie,Kafka																
ทำเอกสารเครื่องมือ Big Data																
นำเสนอเครื่องมือ Big Data มาปรับใช้ในธุรกิจ																
ติดตั้งโปรแกรม Cloudera ขึ้น Cloud A-Host																
ทำเอกสารการติดตั้งCloudera																
Monitoring Report TAT																

3.2 รายละเอียดที่นักศึกษาปฏิบัติในการฝึกงาน

เนื่องจากผู้จัดทำโครงงานได้เข้าร่วมสหกิจในช่วงที่แผนก IBM Channel (BI) เริ่มต้นการรับทำBig Data และได้มี Project Big Data ซึ่งอยู่ในช่วงImplement ดังนั้นผู้จัดทำโครงงานจึงได้รับมอบหมายให้ทำการช่วยนำข้อมูลจากต้นทาง ไปปรับปรุงซึ่งเป็นกระบวนการ ETL ใช้โปรแกรม StreamSets และส่งไปยังปลายทางที่เก็บคือ Hadoop distributed file system (HDFS) และทำการศึกษาเครื่องต่างๆ เพื่อเตรียมความพร้อมในการรับมือกับ requirement ของลูกค้า และต่อยอดศึกษาเครื่องมือ Big Data เพิ่มเติมเพื่อรับมือกับงานของลูกค้าท่านอื่นต่อไป และทำการติดตั้ง Cloudera ซึ่งเป็นตัวจัดการให้เครื่องมือ Big Data ต่างๆ อยู่ในสภาพแวดล้อมเดียวกันและง่ายต่อการจัดการ ขึ้นไปอยู่บน Cloud ของบริษัท

3.3 ขั้นตอนการดำเนินงานที่นักศึกษาปฏิบัติงาน

3.3.1 ศึกษา Business Intelligence และ Big Data

ศึกษา Business Intelligence เพื่อให้เข้าใจการการนำข้อมูลมาใช้ในเชิงธุรกิจ และศึกษา Big Data Concept จะทำให้เข้าใจรวดเร็วยิ่งขึ้น เมื่อต้องปฏิบัติงานจริง

3.3.2 ศึกษากระบวนการ ETL

ได้ศึกษากระบวนการ ETL (Extract transform load) เพื่อที่จะสามารถนำข้อมูลจากต้นทางไปปรับปรุง และส่งไปยังปลาย ผู้จัดทำได้รับมอบหมายให้นำข้อมูลจากต้นทางไปเก็บใน HDFS ซึ่งเป็นที่เก็บข้อมูลแบบกระจาย ที่เป็นรูปแบบการเก็บของ Big Data ด้วยการใช้โปรแกรม StreamSets

3.3.3 ศึกษาการใช้และทำความเข้าใจ StreamSets

ศึกษาการใช้งาน StreamSets ซึ่งเป็นเครื่องมือ ETL เพื่อย้ายข้อมูลของลูกค้า เนื่องจากผู้จัดทำโครงงานไม่เคยเรียนรู้มาก่อน จึงต้องเรียนรู้จากเอกสารและทำแบบฝึกหัดที่พี่พนักงานให้มาเพื่อไปทำการย้ายข้อมูลของลูกค้า

3.3.4 ศึกษาเครื่องมือและทำความเข้าใจ Hue

ศึกษาการใช้งาน Hue ซึ่งเป็น Web UI ที่ให้เครื่องมือของBig Data อื่นๆทำงานบนWeb UI ได้แทนที่การใช้command line และสามารถดูข้อมูลต่างๆใน HDFS ได้ผ่าน Web UI เพื่อให้สามารถใช้เครื่องมือต่างๆของBig Data ได้ง่ายขึ้น

3.3.5 ศึกษาเครื่องมือและทำความเข้าใจ Hive,Impala

ศึกษาการใช้งาน Hive,Impala ซึ่งเป็นเครื่องมือที่ใช่ query ข้อมูลโดยทำบนหน้า Web UI ของ Hue เพื่อให้สามารถตรวจสอบข้อมูลที่อยู่ใน HDFS ว่ามีความถูกต้องหรือไหม

3.3.6 ศึกษาเครื่องมือ Ozzie

ศึกษาการใช้งาน Ozzie ซึ่งเป็นเครื่องมือสำหรับกำหนดการทำงานอัตโนมัติ เพื่อให้สามารถกำหนดเวลาทำงานของเครื่องต่างๆ ได้

3.3.7 ศึกษาระบบ Kafka

ศึกษาระบบ Kafka เพื่อไม่ให้ข้อมูลขาดหายหรือมาไม่สมบูรณ์ระหว่างการรับส่งข้อมูล และนำไปต่อยอดการทำ Real-time processing กับ big data

บทที่ 4
สรุปผลการดำเนินงาน การวิเคราะห์และสรุปผลต่าง ๆ

4.1 ขั้นตอนและผลการดำเนินงาน

ในส่วนของขั้นตอนการดำเนินงานของBig Dataนั้นคือ การดึงข้อมูลจากแหล่งข้อมูลต้นทาง สู่คลังข้อมูลปลายทางนั้นคือHadoop distributed file system (HDFS) ทำการศึกษาหาความรู้ที่จะต้องนำไปใช้ประมาณ 2 สัปดาห์ หลังจากนั้นจึงเริ่มปฏิบัติงานจริง

โดยหลักแล้ว จะแบ่งการทำงานออกเป็น 6 ขั้นตอน คือ

4.1.1 การติดตั้งโปรแกรม Cloudera Manager บน Centos 7

ในส่วนนี้นั้นเป็นขั้นตอนการติดตั้งระบบ hadoop เพื่อให้ทำงานอย่างมีประสิทธิภาพจึงใช้ Cloudera Platform ซึ่งเป็นตัวทำให้เครื่องมือต่างๆ อยู่ในสภาพแวดล้อมเดียวกันและง่ายต่อการจัดการ

4.1.1.1 การเตรียมไฟล์ และนำไปเก็บใน Directory /opt

1) CDH 5 parcel และ SHA

CDH หรือ Cloudera Distributed Hadoop เป็น Open-Source Platform ที่ provide Big Data Tools เอาไว้มี 2 ไฟล์ด้วยกัน CDH.parcel และ CDH.parcel.sha1 ดาวน์โหลดตามลิ้งนี้ https://archive.cloudera.com/cdh5/parcels/5.15.0/ ทำการเลือกversion OS และโหลด เมื่อโหลดเสร็จเปลี่ยนไฟล์นามสกุล.sha1 เป็น .sha

2) CM5 repo-tarball

เป็นไฟล์หลักของ Cloudera Manager สำหรับติดตั้ง Cloudera Agent ดาวน์โหลดตามลิ้งนี้ https://archive.cloudera.com/cm5/repo-as-tarball/5.15.0/

3) Cloudera package

เป็น package ของ Cloudera ดาวน์โหลดได้ตามลิ้งนี้ https://archive.cloudera.com/cm5/redhat/7/x86_64/cm/5.15.0/RPMS/x86_64/ เลือกโหลด 4 package ด้วยกันได้แก่

- cloudera-manager-server
- cloudera-manager-daemons
- cloudera-manager-server-db

- enterprise-debuginfo

4) JDK 7 or 8

JDK หรือ Java SE Development Kit เป็นpackage ที่ช่วยในการติดตั้ง Cloudera agent ดาวน์โหลดได้ตามลิ้งนี้ http://www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase7-521261.html (jdk v.7) และ http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html (jdk v.8)

5) mysql-connector-java

เป็นตัวที่ทำให้ Database ของเครื่องติดต่อกับ java ได้ https://cdn.mysql.com//Downloads/Connector-J/mysql-connector-java-5.1.46.tar.gz

4.1.1.2 ติดตั้ง Package ที่ Cloudera ต้องการ

- python (2.7 required for CDH 5)
- mod_ssl
- MySQL-python
- cyrus-sasl-gssapi
- python-psycopg2
- openssl-devel
- postgresql-server >=8.4
- httpd
- mariadb-server (คือฐานข้อมูลที่เลือกใช้)
- redhat-lsb-core

4.1.1.3 การตั้งค่าเบื้องต้น

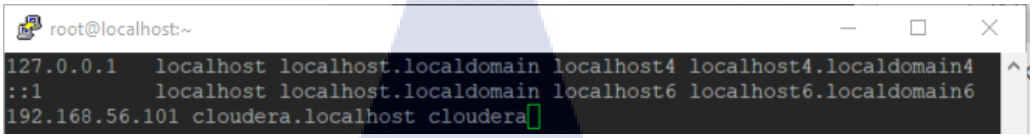
1) ตั้งค่า hosts

vi /etc/hosts

รูปแบบคือ Ip localhost hostname

ยกตัวอย่าง.

192.168.56.101 cloudera.localhost cloudera

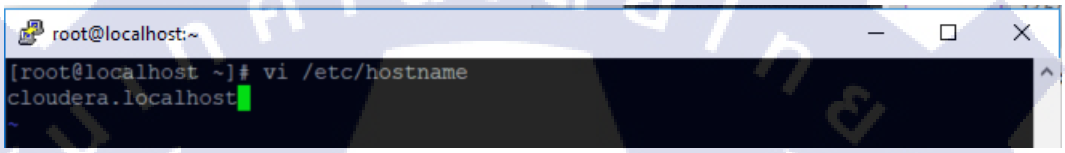


รูปที่ 4.1 ตัวอย่างการตั้งค่า

2) ตั้งค่า hostname

vi /etc/hostname

รูปแบบคือ FQDN (fully qualified domain name)



รูปที่ 4.2 ตัวอย่างการตั้งค่า

3) ปิด Firewall

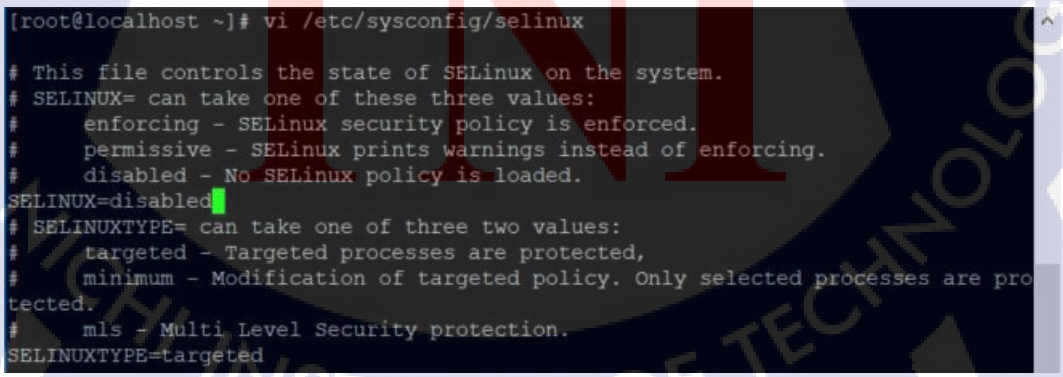
systemctl stop firewalld

systemctl disable firewalld

4) ตั้งค่า selinux เป็น disabled

Selinux เป็นตัวดักจับ package ชนิดหนึ่งของค่าย CentOS

ทำการเปลี่ยนดังรูปที่ 4.3

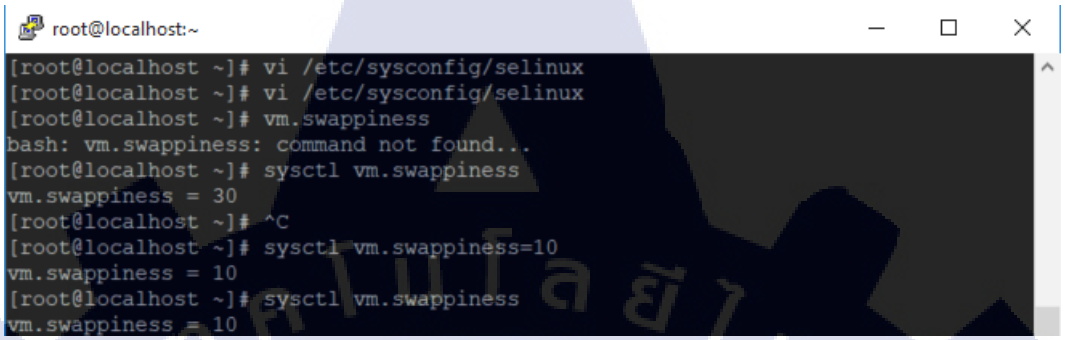


รูปที่ 4.3 การเปลี่ยน selinux เป็น Disabled

5) ตั้งค่า swappiness ให้น้อยกว่าหรือเท่ากับ 10

ถ้ามากกว่า 10 ต้องทำให้น้อยกว่าหรือเท่ากับ 10 ในที่นี้ผู้จัดทำการตั้งค่าไว้ที่ 10

#sysctl vm.swappiness=10



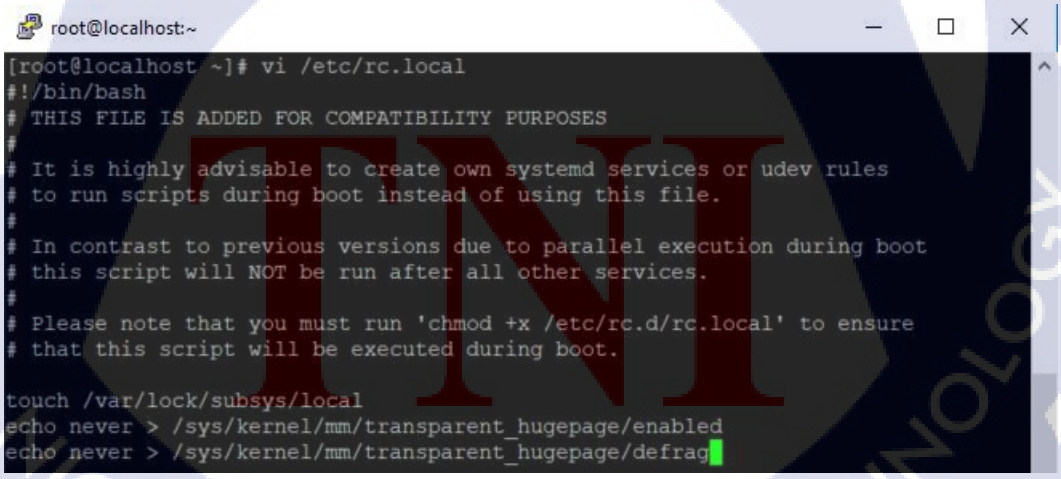
รูปที่ 4.4 การตั้งค่า swappiness

6) ตั้งค่า rc.local

vi /etc/rc.local

echo never > /sys/kernel/mm/transparent_hugepage/enabled

echo never > /sys/kernel/mm/transparent_hugepage/defrag



รูปที่ 4.5 การตั้งค่า rc.local

เมื่อตั้งค่าเสร็จให้ออกและecho อีกครั้ง

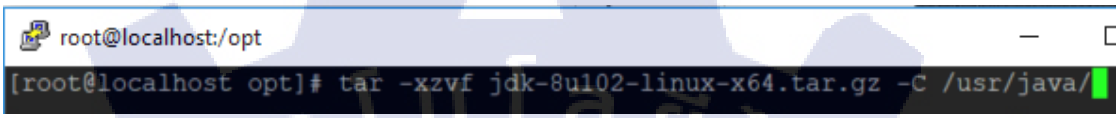
echo never > /sys/kernel/mm/transparent_hugepage/enabled

echo never > /sys/kernel/mm/transparent_hugepage/defrag

7) ติดตั้ง package JDK

cd /opt

tar -xzvf /path/to/JDKversion -C /usr/java/



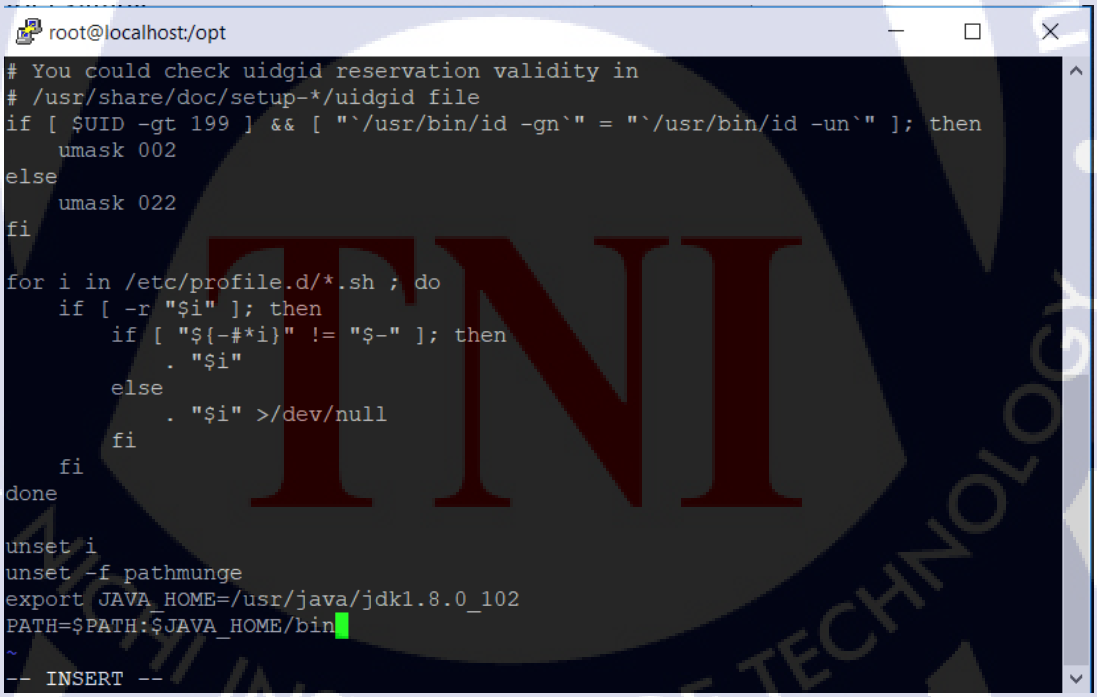
รูปที่ 4.6 การใช้คำสั่งtar ไฟล์ไปยัง directoryที่ต้องการ

vi /etc/profile

เพิ่มลงในprofile

export JAVA_HOME=/usr/java/jdk1.8.0_102

PATH=$PATH:$JAVA_HOME/bin



รูปที่ 4.7 การใส่ path ให้กับที่อยู่ของJDK

เมื่อเสร็จแล้วให้ทำการreboot เครื่อง

reboot

8) ติดตั้ง package cloudera

rpm -ivh cloudera-manager-daemons

rpm -ivh cloudera-manager-server

rpm -ivh cloudera-manager-server-db

rpm -ivh enterprise-debuginfo

9) ย้ายไฟล์ parcel CDH ไปยัง /opt/cloudera/parcel-repo

mv /opt/CDH.parcel /opt/cloudera/parcel-repo

mv /opt/CDH.parcel.sha /opt/cloudera/parcel-repo

4.1.1.4 ทำการ tar ไฟล์ CM ไปยัง /var/www/html/repo

cd /opt

mkdir -p /var/www/html/repo

tar -xzvf cm5.15.0-centos7.tar.gz -C /var/www/html/repo

ตรวจสอบในdirectory /etc/yum.repos.d ว่ามีไฟล์repoอื่นๆหรือไหมถ้ามีทำให้ไม่มีแล้ว copy repoของ CM มาใส่ในdirectoryนี้

cp /var/www/html/repo/cm/cloudera-cm.repo /etc/yum.repos.d/cloudera-manager.repo

ทำการกำหนดให้repoเรียกใช้localhost

cd /etc/yum.repos.d

vi cloudera-manager.repo

ทำการแก้ไข path

baseurl = http://localhost/repo/cm/5

gpgkey = http://localhost/repo/cm/RPM-GPG-KEY-cloudera และ start service httpd

#systemctl start service http

root@cloudera:/etc/yum.repos.d

```
[cloudera-cm]
# Packages for Cloudera's Distribution for cm, Version 5, on RedHat     or CentOS 7 x86_64
name=Cloudera's Distribution for cm, Version 5
baseurl=http://cloudera.localhost/repo/cm/5
gpgkey = http://cloudera.localhost/repo/cm/RPM-GPG-KEY-cloudera
gpgcheck = 1
~
```

รูปที่ 4.8 ตัวอย่างการแก้ path

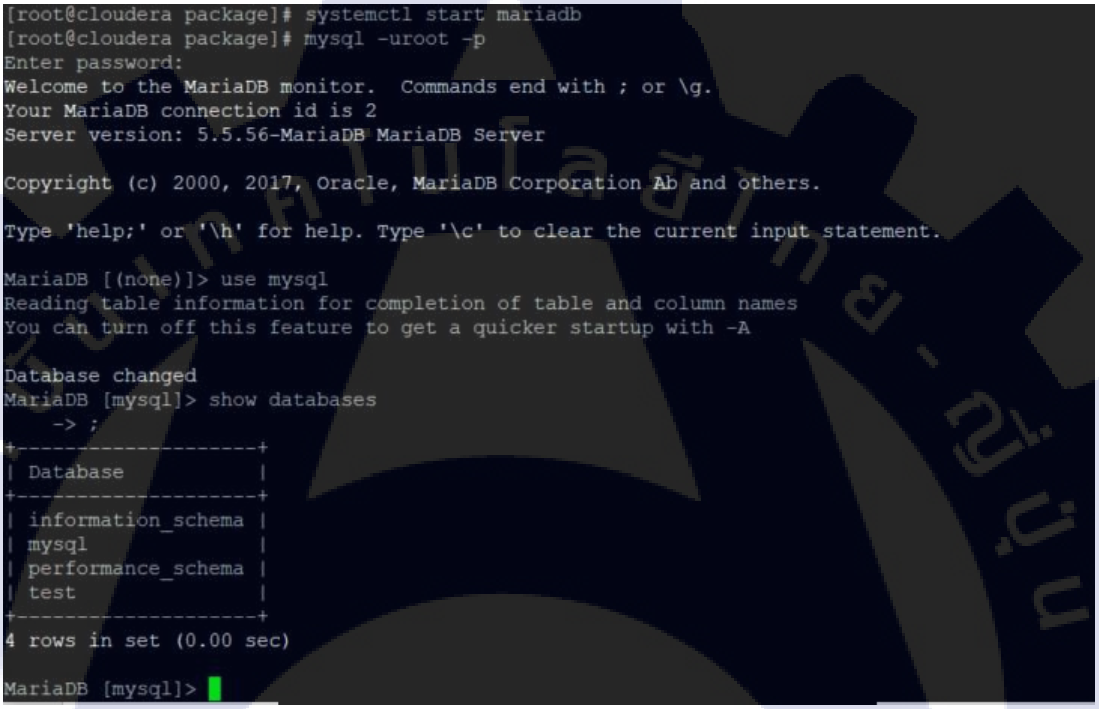
4.1.1.5 ตั้งค่า MariaDB

systemctl start mariadb

mysql -uroot -p

Enter password: [Enter] ผ่านได้เลย

MariaDB [(none)] > use mysql



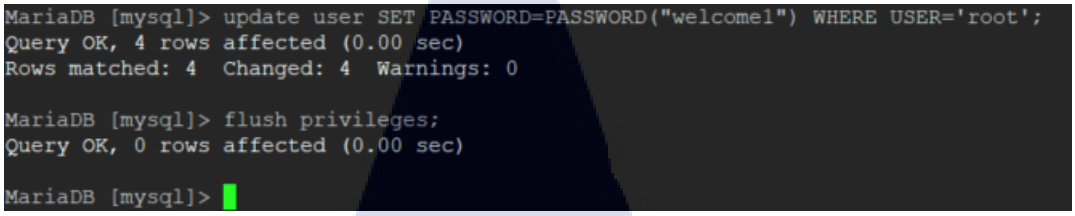
รูปที่ 4.9 การเข้า mariadb ด้วย command line

สามารถเปลี่ยนรหัสของ Root ด้วย

MariaDB [mysql]> update user SET PASSWORD=PASSWORD(“welcome1”) WHEREUSER=’root’;

ให้แน่ใจว่าได้บันทึกการเปลี่ยนแปลงโดยการออกคำสั่งต่อไปนี้

MariaDB [mysql]> flush privileges;



รูปที่ 4.10 ตัวอย่างการเปลี่ยนรหัสผ่าน

ทำการกำหนดสิทธิ์ให้กับ Root

MariaDB [(none)]> grant all privileges on *.* to 'root'@'cloudera.localhost' identified by 'welcome1' with grant option;

MariaDB [(none)]> flush privileges;

ทำการติดตั้ง MySQL JDBC driver

cd /opt

tar -xzvf mysql-connector-java-5.1.46.tar.gz

cp mysql-connector-java-5.1.46-bin.jar /usr/share/java/mysql-connector-java.jar

ถ้าไม่มีdirectory ให้ใช้คำสั่งนี้

mkdir -p /usr/share/java/

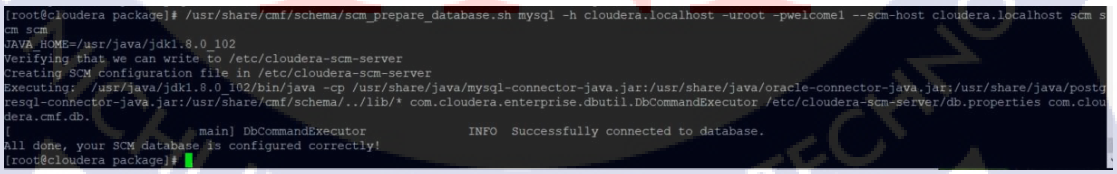
ทำการเตรียมฐานข้อมูลให้กับ Cloudera Manager Server External Database

Syntax

/usr/share/cmf/schema/scm_prepare_database.sh database-type [options] database-name username password

ตัวอย่าง

/usr/share/cmf/schema/scm_prepare_database.sh mysql -h cloudera.localhost -uroot -pwelcome1 --scm-host cloudera.localhost scm scm scm



รูปที่ 4.11 ตัวอย่าง syntax

4.1.1.6 การติดตั้ง Cloudera agent และ CDH

เปิด service cloudera-server

systemctl start cloudera-scm-server

ไปยัง web browser เพื่อเข้าเปิด web UI ของ cloudera manager

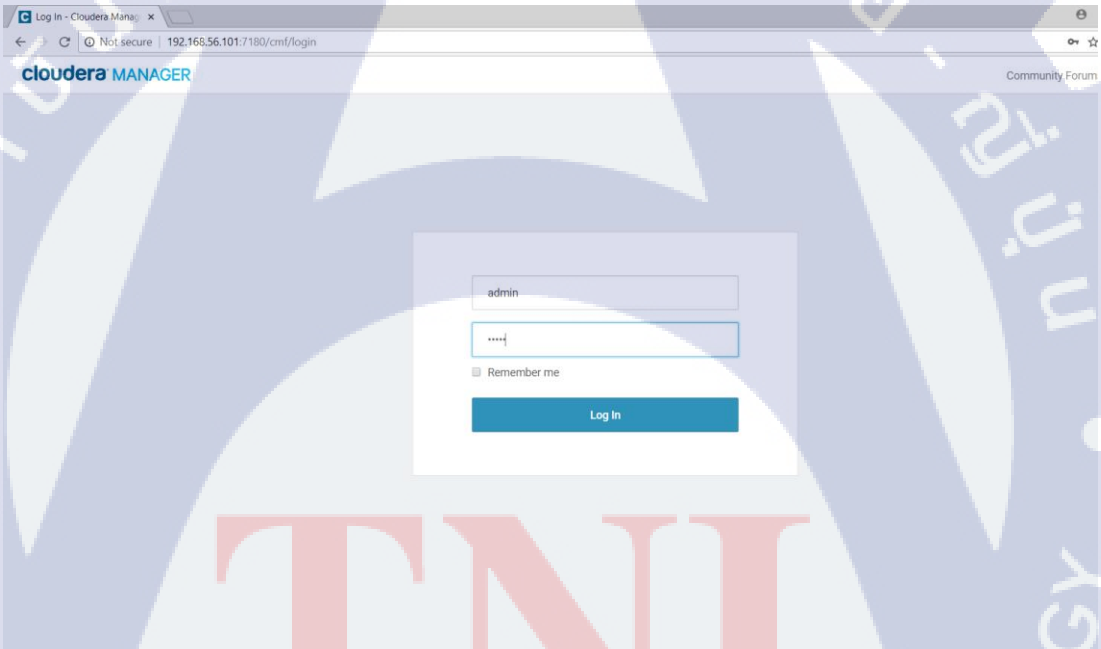
http://localhost:7180 or 127.0.0.1:7180

ผู้จัดทำใช้

http://cloudera.localhost:7180 or 192.168.56.101:7180

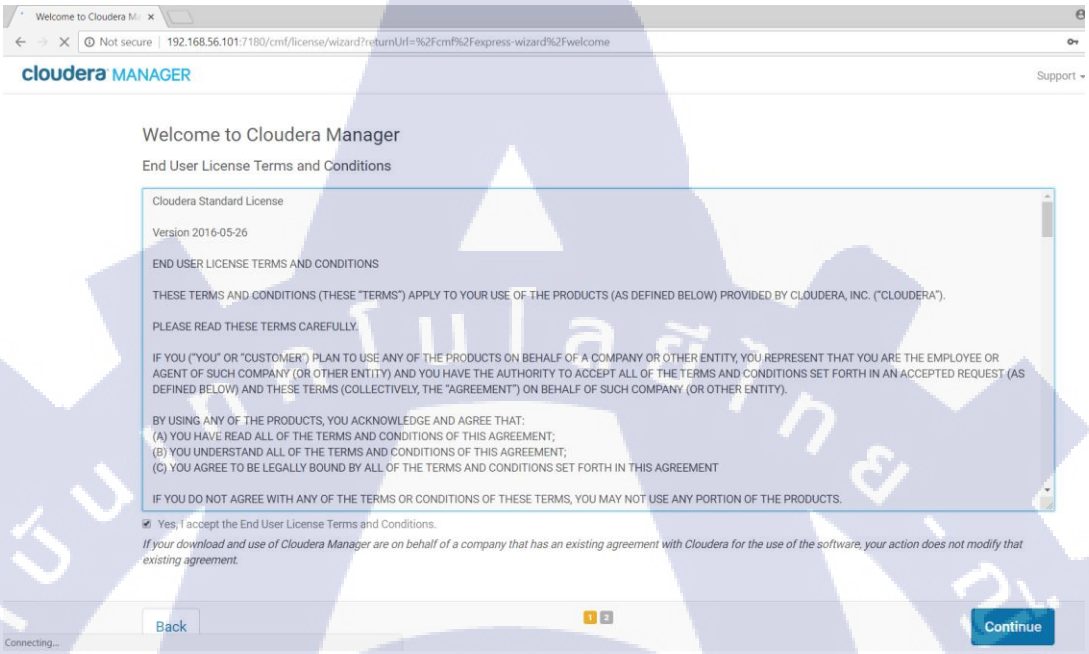
User: admin

password: admin



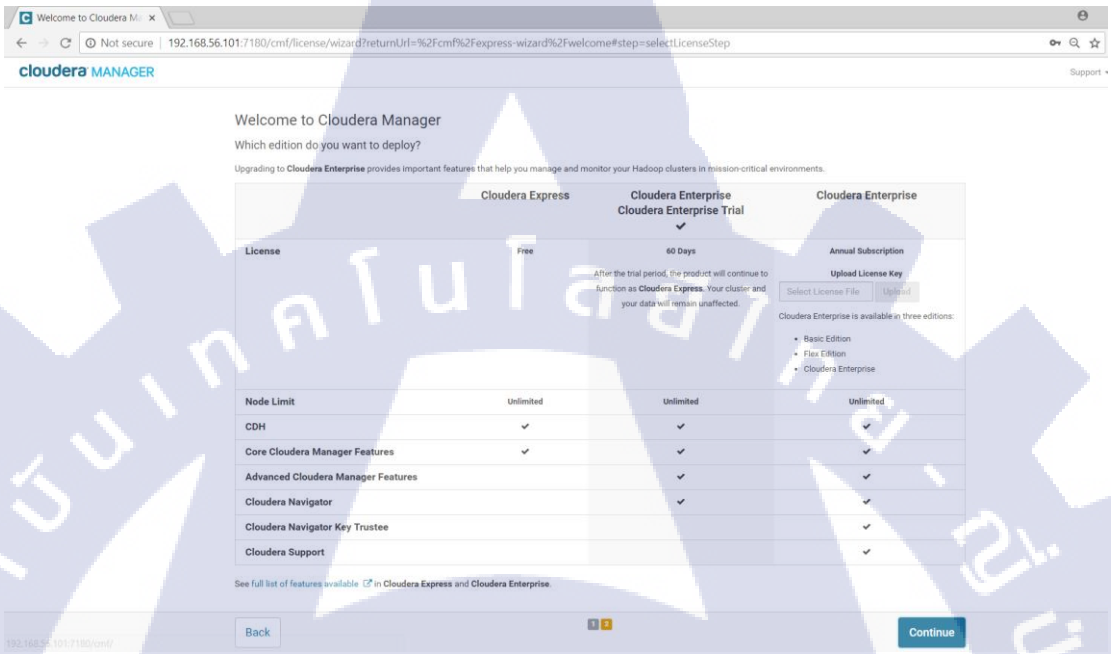
รูปที่ 4.12 หน้าWeb UI cloudera Manager

เมื่อเข้าไปจะเจอหน้าให้อ่าน license ดังรูปที่ 4.13 ให้เลือก Yes,I accept และกด continue



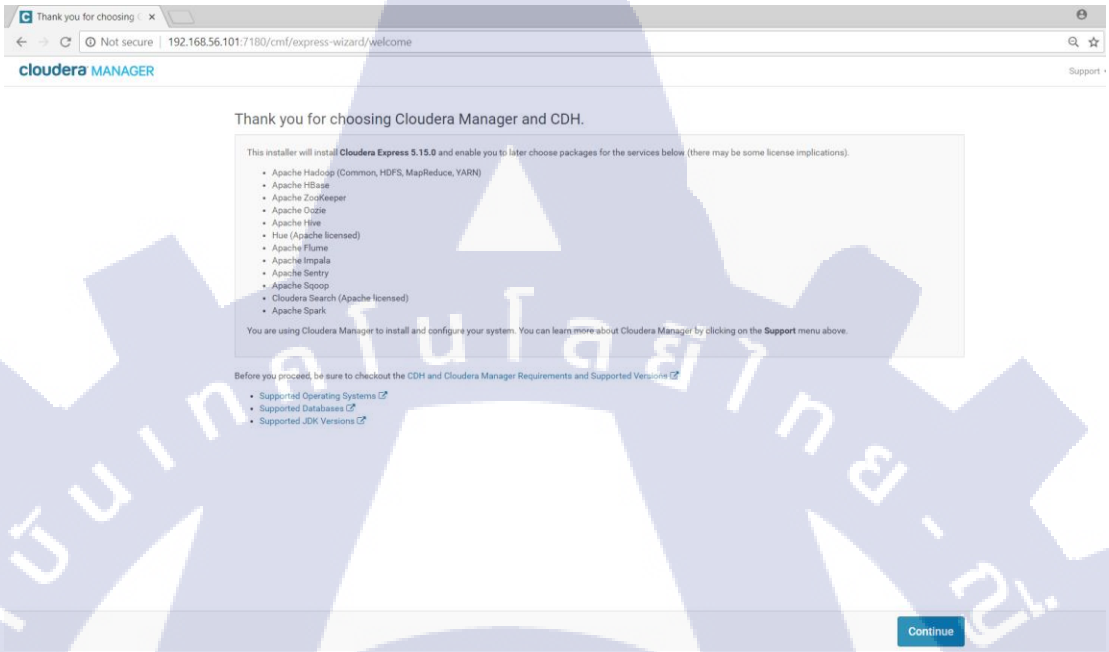
รูปที่ 4.13 User License

จะเจอหน้าให้เลือกรูปแบบการใช้งานดังรูปที่ 4.14 และเลือกแบบ Cloudera Express ซึ่งเป็นแบบ Free



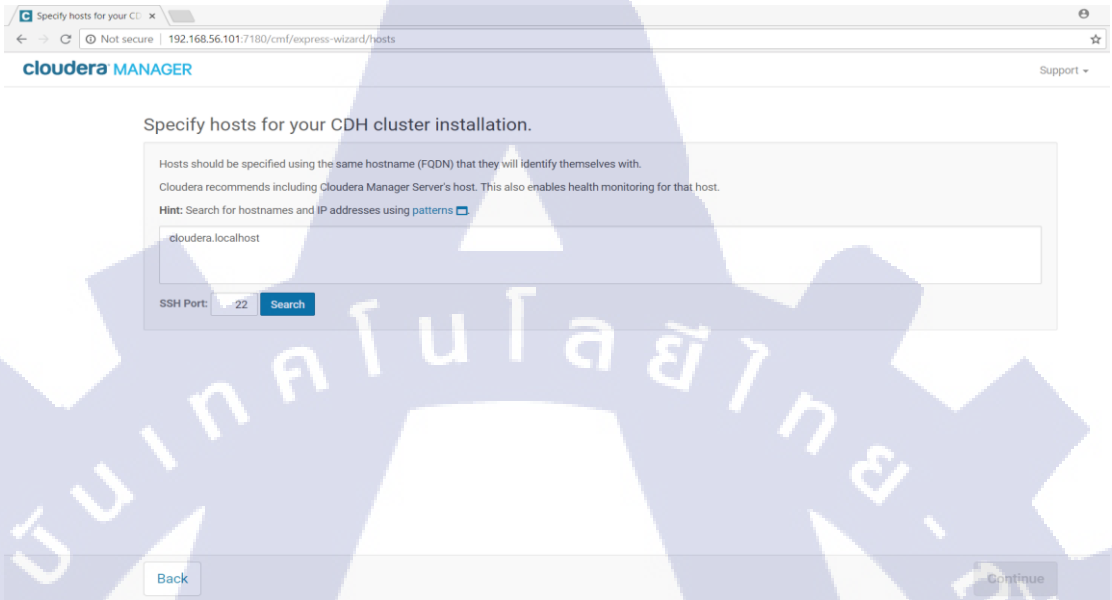
รูปที่ 4.14 หน้าเลือกรูปแบบการทำงาน

จะเจอหน้า thank you for choosing Cloudera Manager and CDH ดังรูปที่ 4.15ให้กด continue



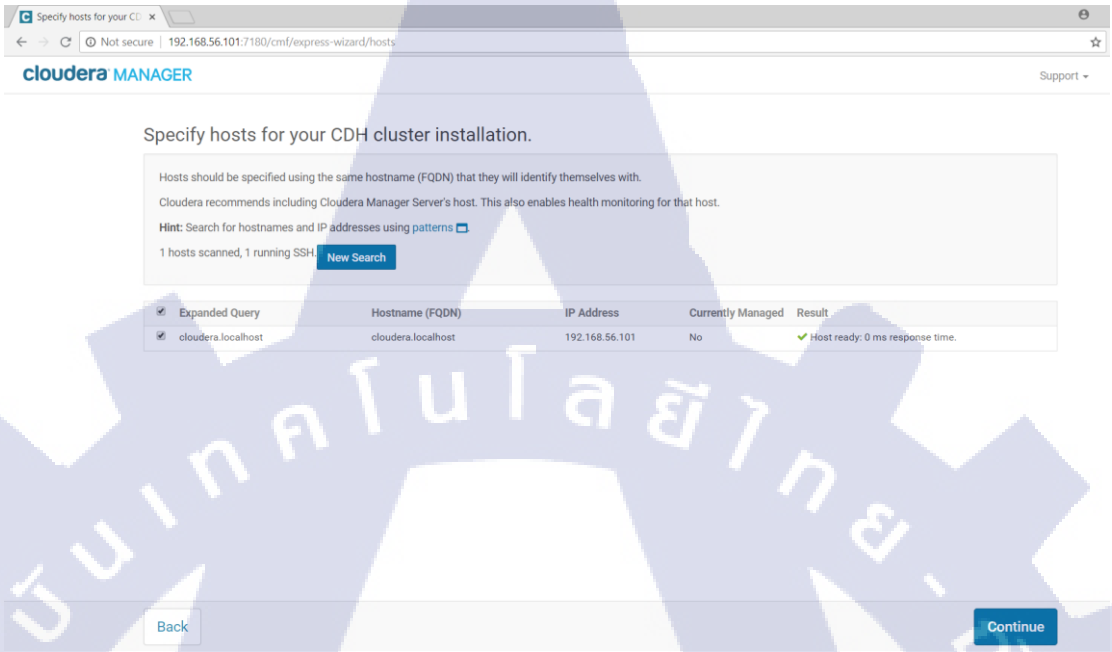
รูปที่ 4.15 thank you for choosing Cloudera Manager and CDH

เมื่อถึงหน้านี้ให้ใส่ hosts ที่เราจะใช้เป็นเครื่องcluster ในการทำงานของCloudera Manager ทั้งหมด จะใส่เป็นhostnames หรือ IP addresses ก็ได้ดังรูปที่ 4.16 ตั้งค่า hosts



รูปที่ 4.16 ตั้งค่า hosts

ในที่นี้ผู้จัดทำใส่ Cloudera.localhost และกดSearch จะได้ดังรูปที่ 4.17และให้กดContinue



รูปที่ 4.17 ตั้งค่า Host เมื่อกดsearch

ต่อมาหน้า select repository ให้ตั้งค่าตามนี้

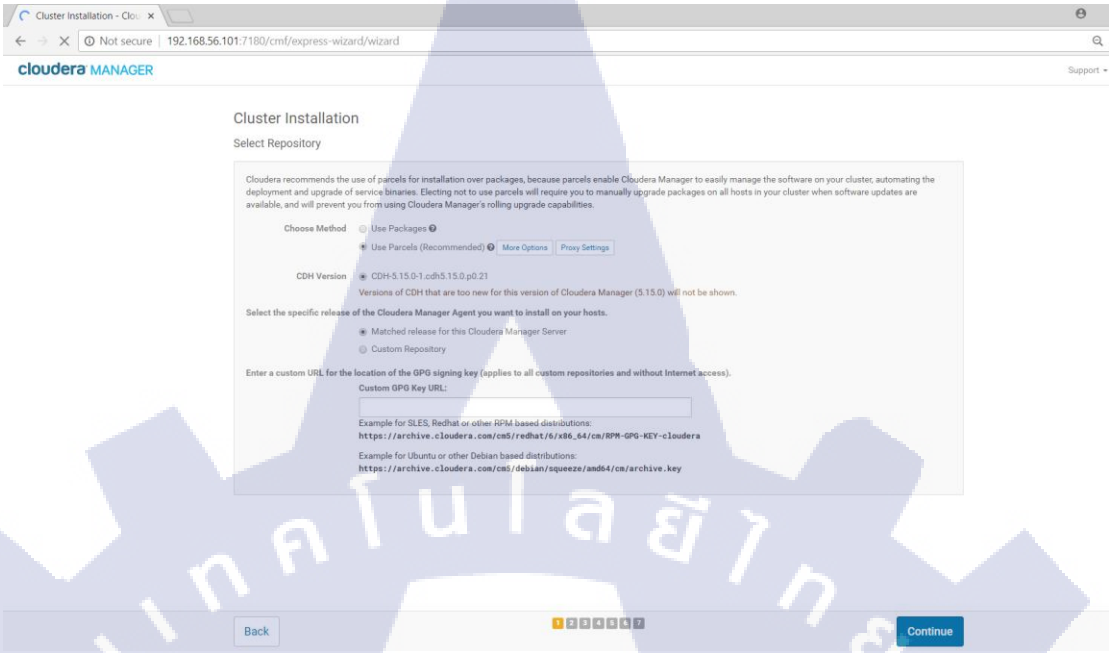
Choose method ให้เลือก Use Parcels

Select the specific release of the Cloudera Manager Agent you want to install on your hosts. ให้เลือก Custom Repository

Custom Repository ของ Cloudera Manager Agent ให้ใส่ http://localhost/repo/cm/5 ในที่นี้ผู้จัดทำใส่ http://192.168.56.101/repo/cm/5

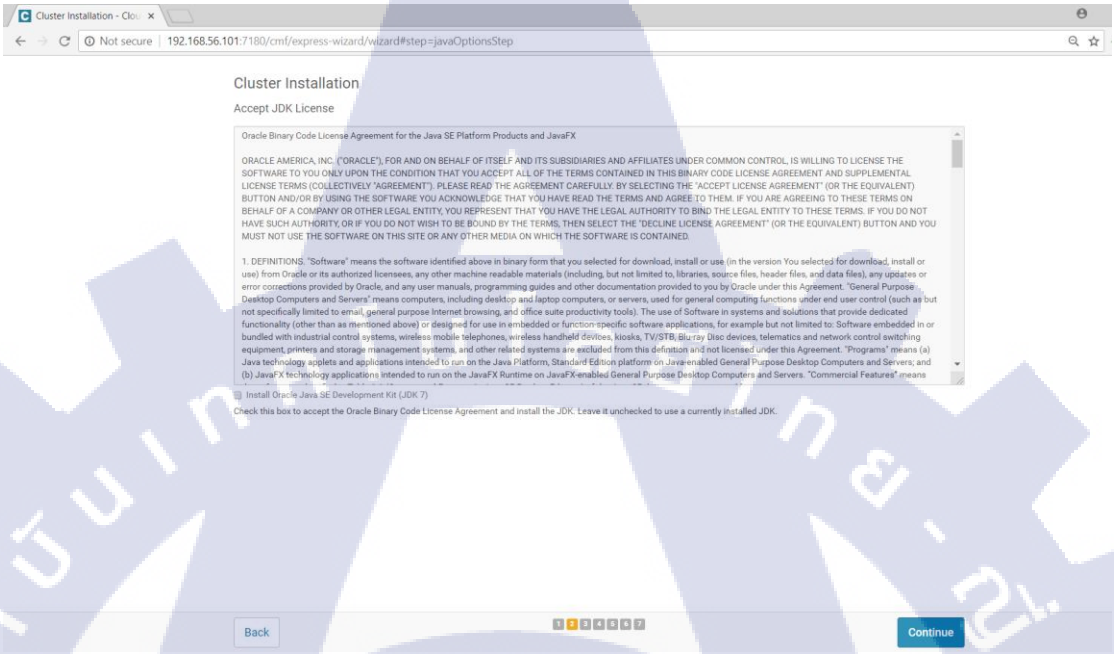
Custom Repository ของ GPG Key URL ให้ใส่ http://localhost/repo/cm/RPM-GPG-KEY-cloudera

ในที่นี้ผู้จัดทำใส่ http://192.168.56.101/repo/cm/RPM-GPG-KEY-cloudera และกด Continue ดังรูปที่ 4.18



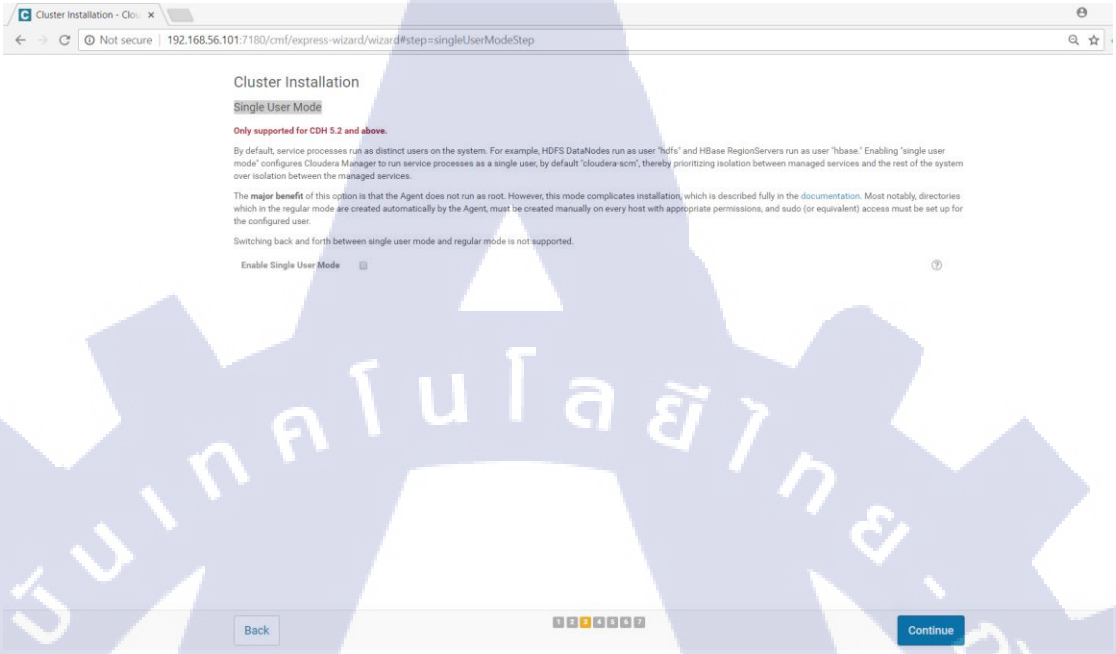
รูปที่ 4.18 การเลือก Repository

ต่อมาจะเจอหน้า Accept JDK License ดังรูปที่ 4.19เนื่องจากผู้จัดทำได้ให้ทำการติดตั้งแล้ว จึงไม่ต้องติดตั้งให้กดอีกให้กด Continue



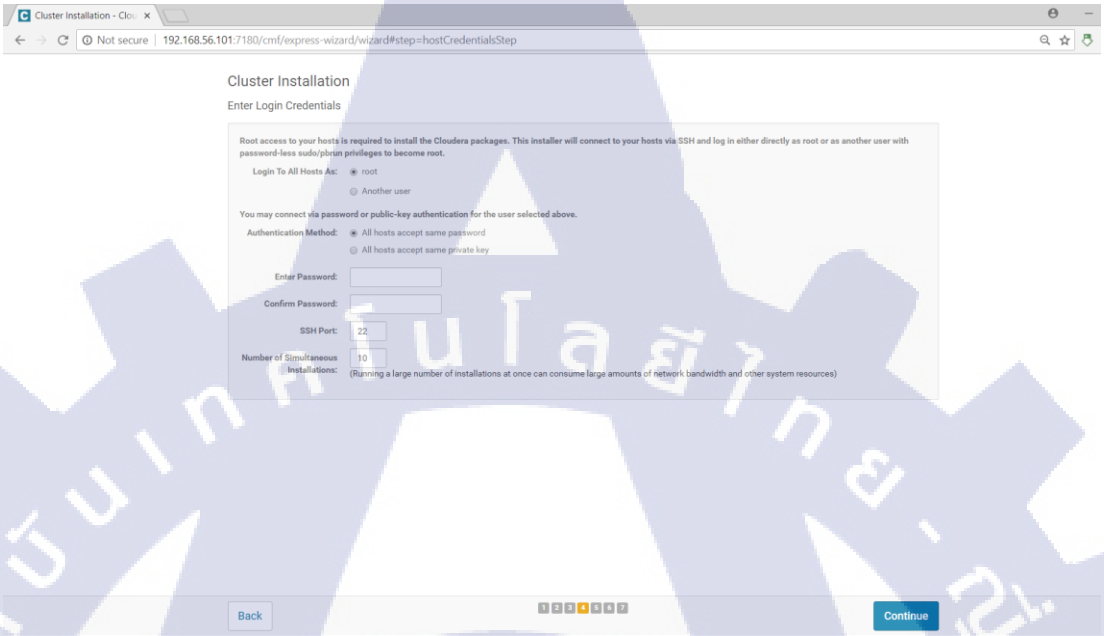
รูปที่ 4.19 หน้า Accept JDk License

ต่อมาจะเจอหน้า Singel User Mode ดังรูปที่ 4.20 ให้กดContinue



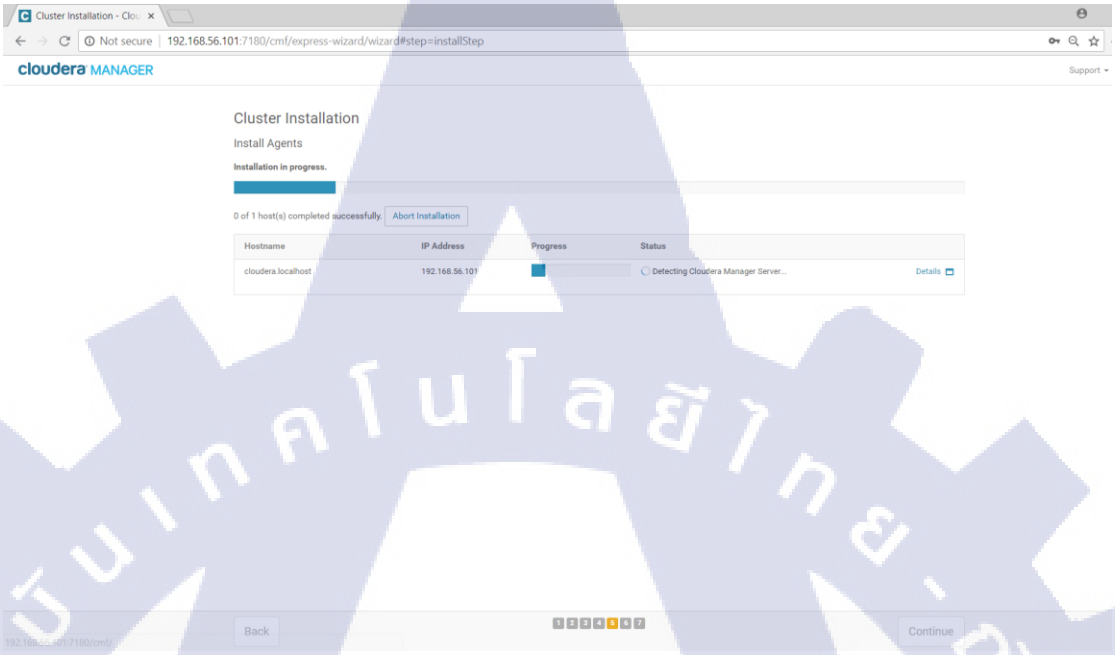
รูปที่ 4.20 หน้า Single user mode

ต่อมาหน้า Enter Login Credentials ดังผู้จัดทำใช่Root ในการlogin จึงทำการใส่แค่password SSH Port:22 number of Simultaneous : 10



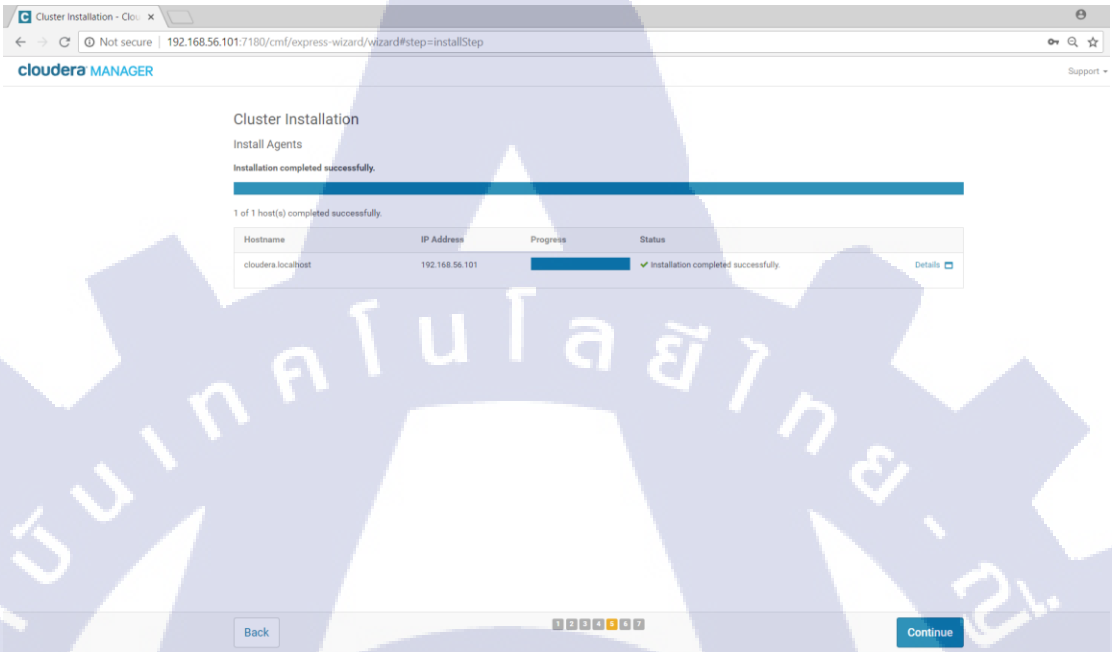
รูปที่ 4.21 หน้า Enter Login Credentials

กด Continue เพื่อไปยังหน้า Install agent ดังรูปที่ 4.22



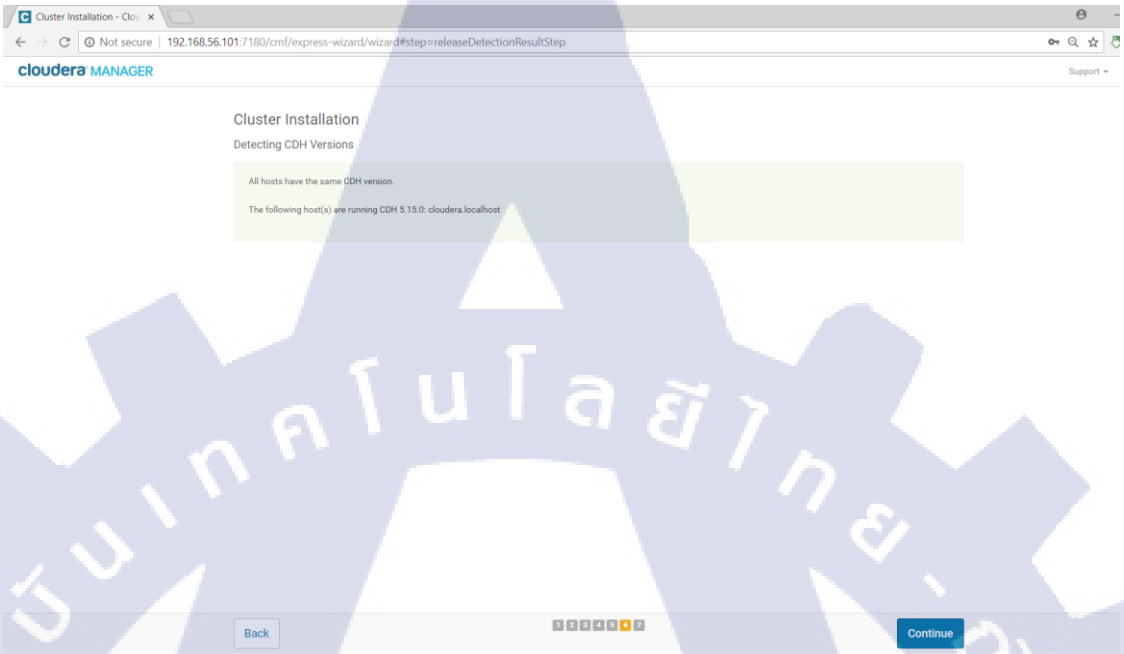
รูปที่ 4.22 หน้า Install agent

เมื่อติดตั้งเสร็จจะเป็นดังรูปที่ 4.23 สามารถกดดูข้อมูลการติดตั้งได้โดยกดปุ่ม Detail และกด Continue เพื่อไปยังหน้าต่อไป



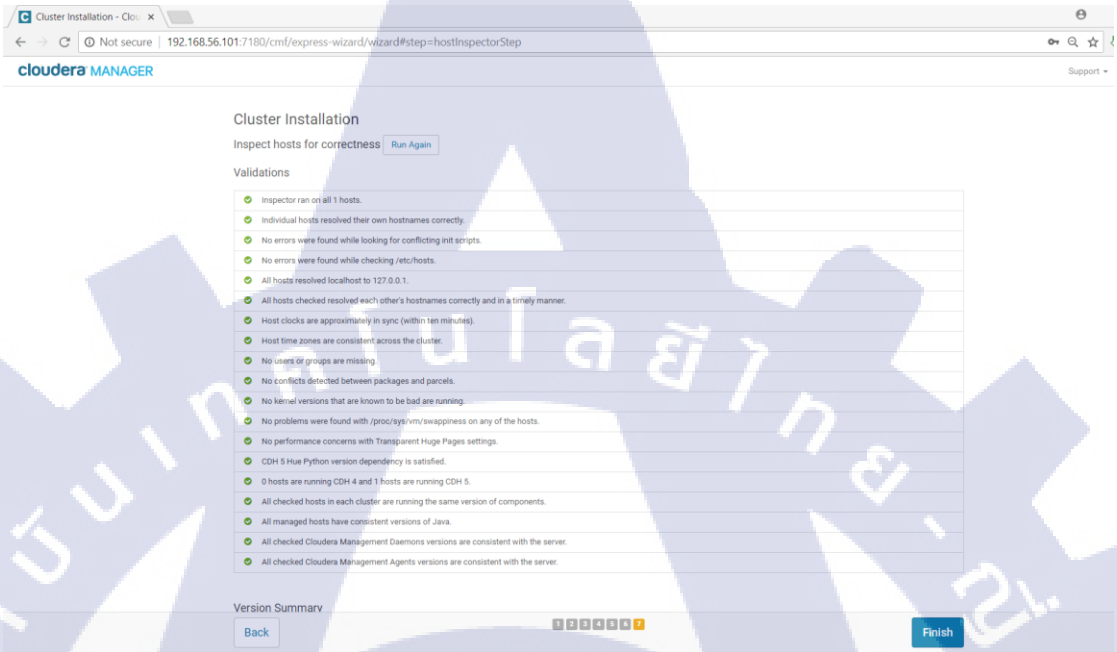
รูปที่ 4.23 Install agent success

จะเจอหน้า Detecting CDH Versions ดังรูปที่ 4.24 ให้กด Continue เพื่อไปยังหน้าต่อไป



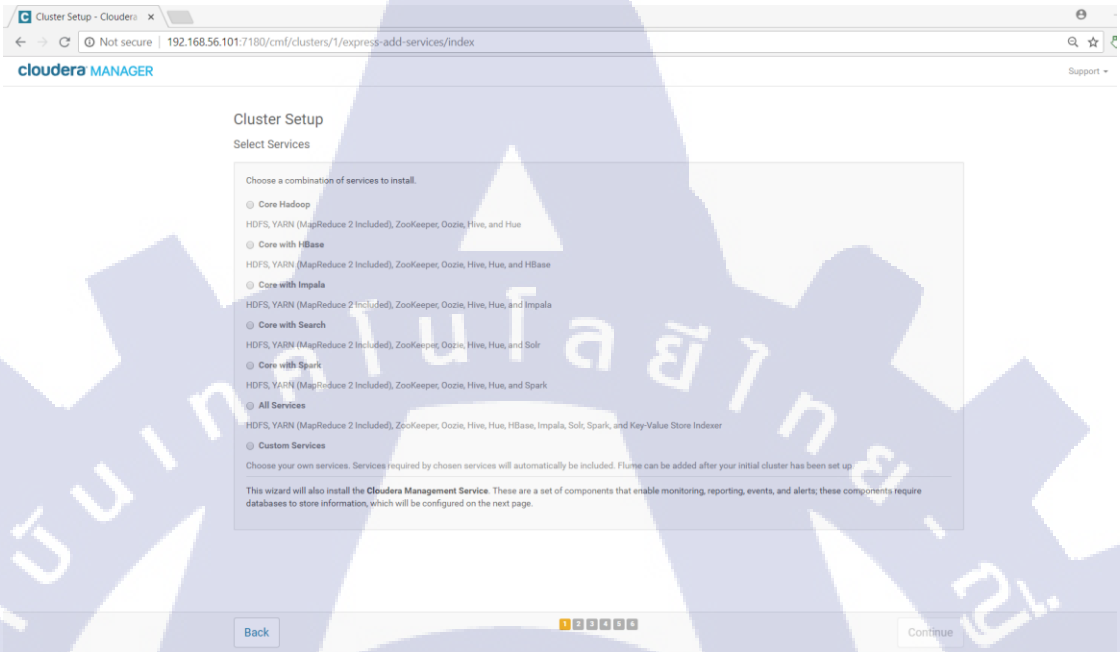
รูปที่ 4.24 หน้า Detecting CDH Versions

ต่อมาจะเป็นหน้า Inspect hosts for correctness ดังรูปที่ 4.25 ให้กด Finish เพื่อไปยังหน้าต่อไป



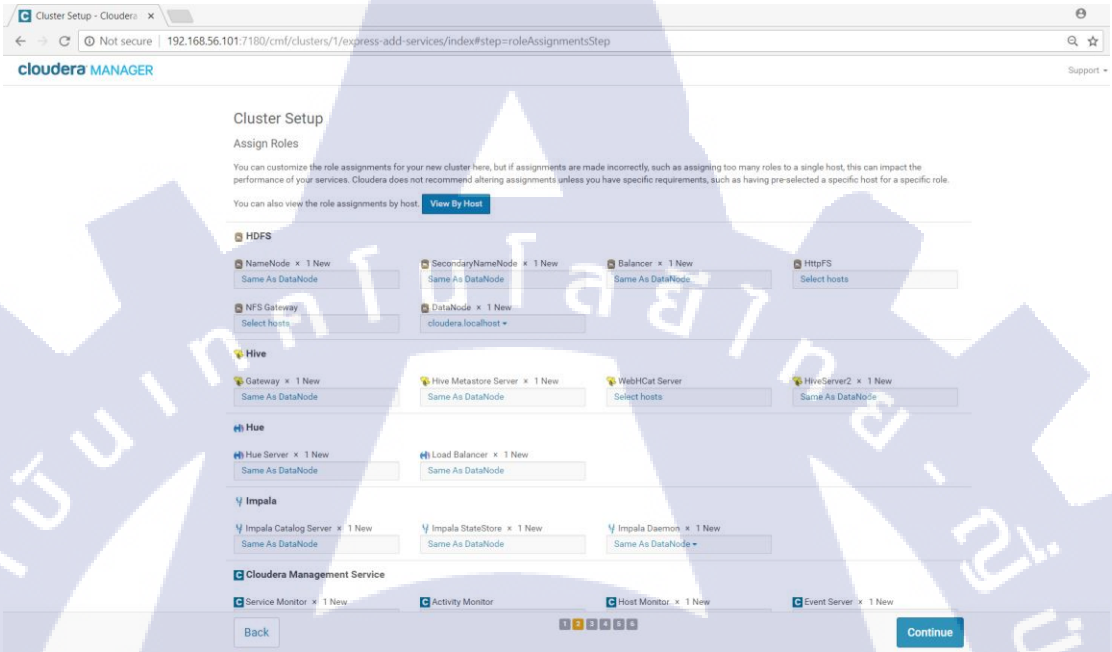
รูปที่ 4.25 หน้า Inspect hosts for correctness

ต่อมาจะเจอหน้า Cluster Setup Select Service ดังรูปที่ 4.26 ในที่นี้ผู้จัดทำเลือก core with Impala และกด Comtinue เพื่อไปยังหน้าถัดไป



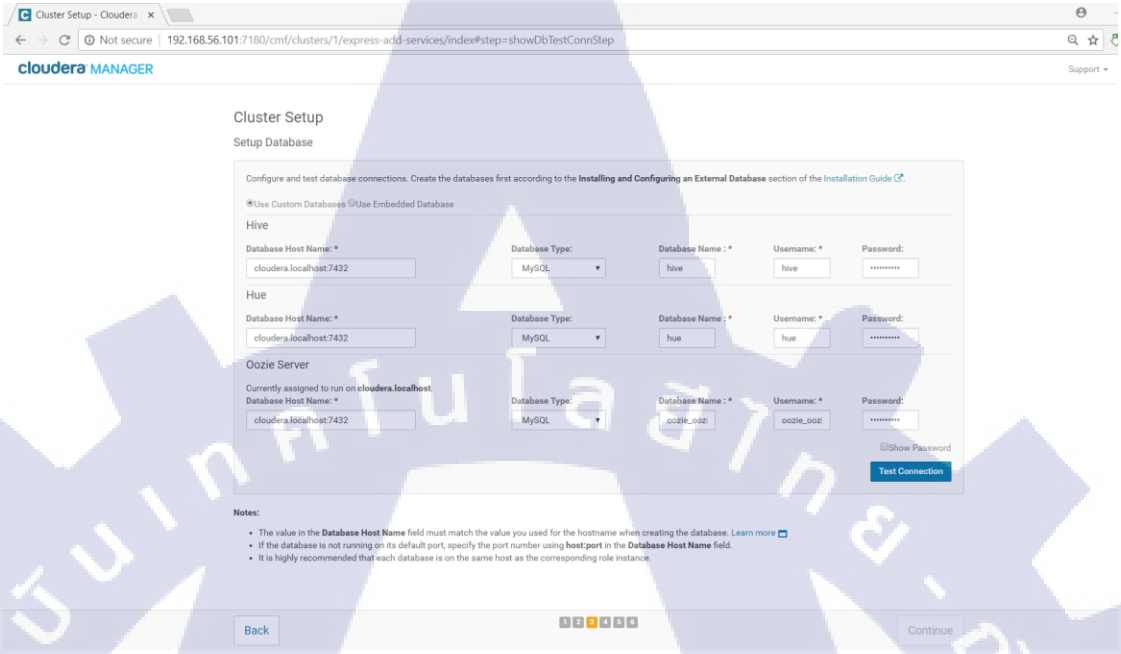
รูปที่ 4.26 หน้า Select Service

ต่อมาหน้า Assign Roles ดังรูปที่ 4.27 เป็นหน้าให้เลือกว่าจะให้การทำงานแบบไหนติดตั้งบนเครื่องไหน แต่ในที่นี้ผู้จัดทำมีเครื่องทำงานเพียงเครื่องเดียวค่าที่ตั้งมาเป็นDefualt ของ cloudera จึงไม่ได้ตั้งค่าใดๆเพิ่มเติมและกด Continue เพื่อไปยังหน้าต่อไป



รูปที่ 4.27 หน้า assign roles

ต่อมาจะเป็น Setup Database ดังรูปที่ 4.28



รูปที่ 4.28 หน้า Setup Database

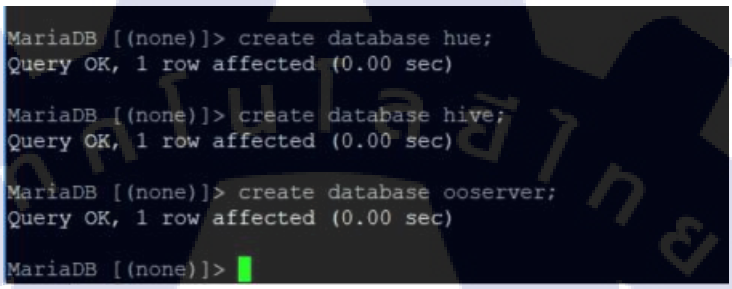
ก่อนอื่นจะต้องทำการสร้าง Database ตามที่ cloudera require ไปที่MariaDB และสร้าง database ใหม่ขึ้นมา

mysql -uroot -p

MariaDB [(none)]> create database hue;

MariaDB [(none)]> create database hive;

MariaDB [(none)]> create database ooserver;



รูปที่ 4.29 สร้าง database

MariaDB [(none)]> flush privileges;

MariaDB [(none)]> exit;

เพื่อออกจากMariaDB

กลับไปยังCloudera Manager

เลือก Use Custom Databases

Hive

Database Host Name: cloudera.localhost,Database type: MySQL,Database Name: hive Username: root,Password: welcome1

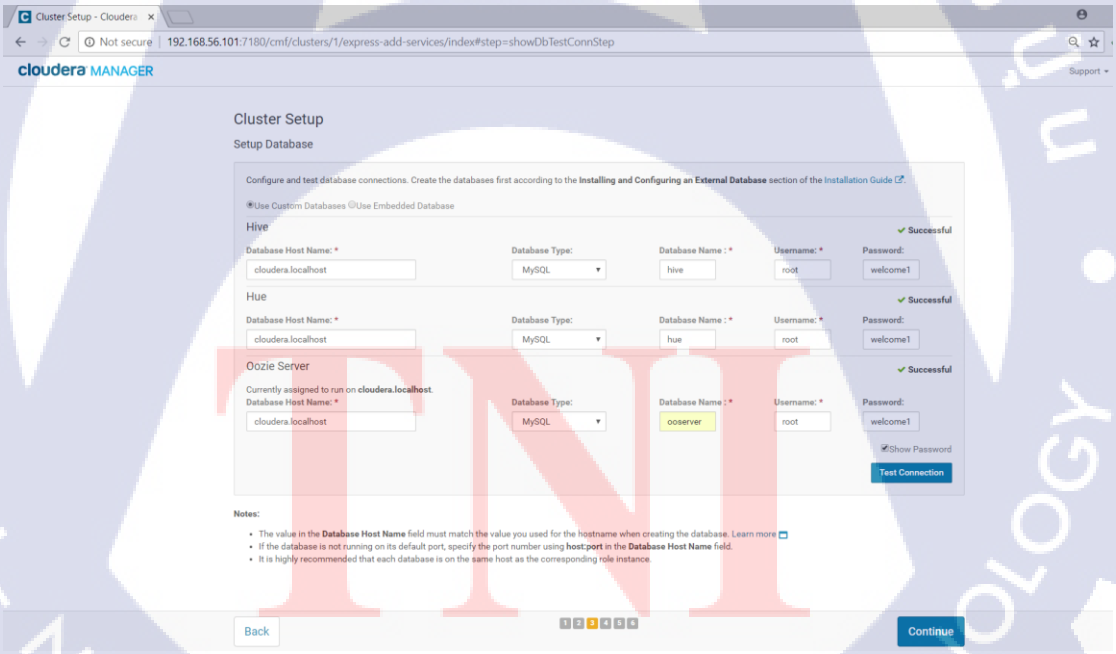
Hue

Database Host Name: cloudera.localhost, Database type: MySQL, Database Name: Hue Username: root, Password: welcome1

Oozie_server

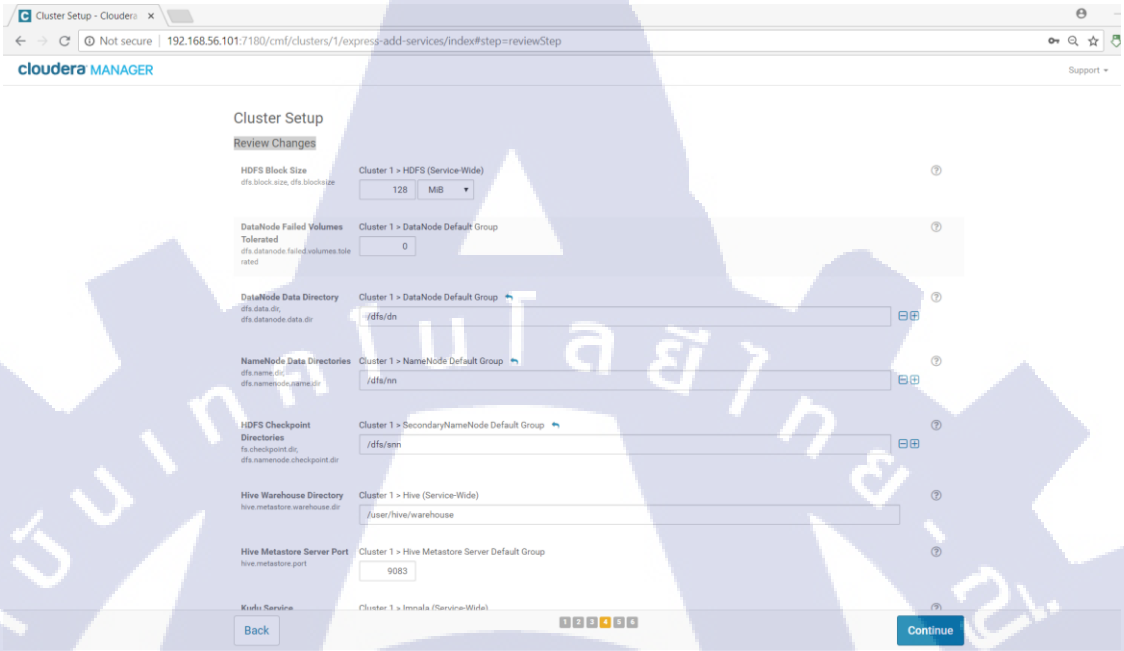
Database Host Name: cloudera.localhost, Database type: MySQL,database Name: ooserver Username: root, Password: welcome1

แล้วกด Test Connection ดังรูปที่ 4.30 เมอ Successful ทั้งหมดให้กด Continue เพื่อไปยังหน้าต่อไป



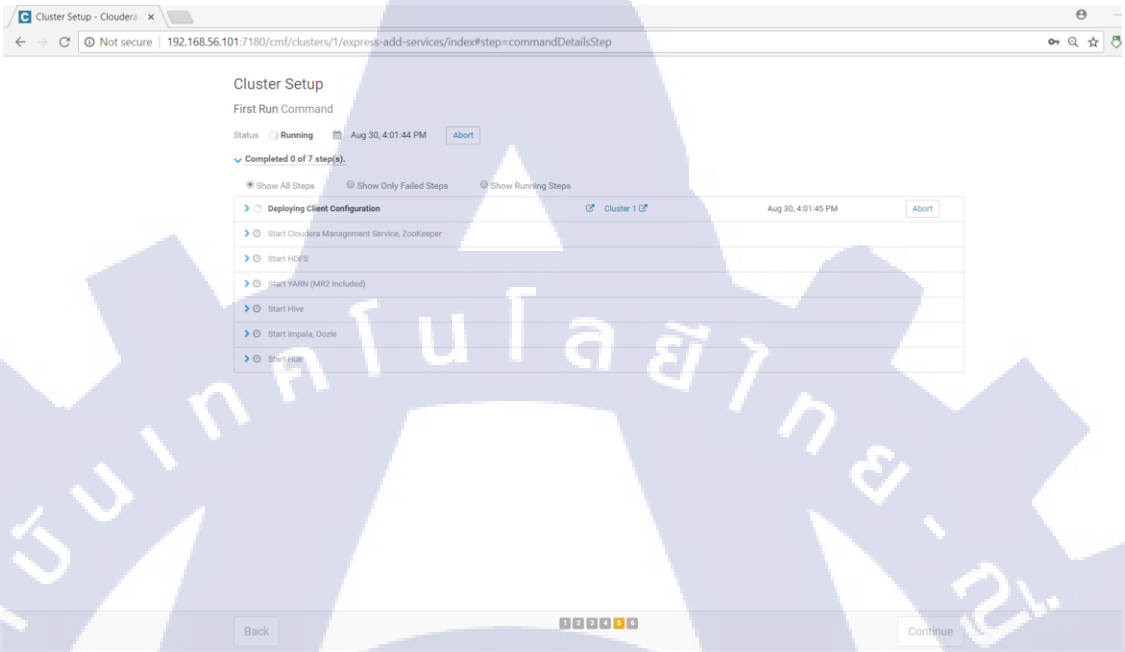
รูปที่ 4.30 Test connection success

ต่อมาจะเป็นหน้า Review Changes ดังรูปที่ 4.31 การตั้งค่าทั้งหมดจะเป็น Default จึงไม่ต้องตั้งค่าอะไรเพิ่มเติมและกด Continue



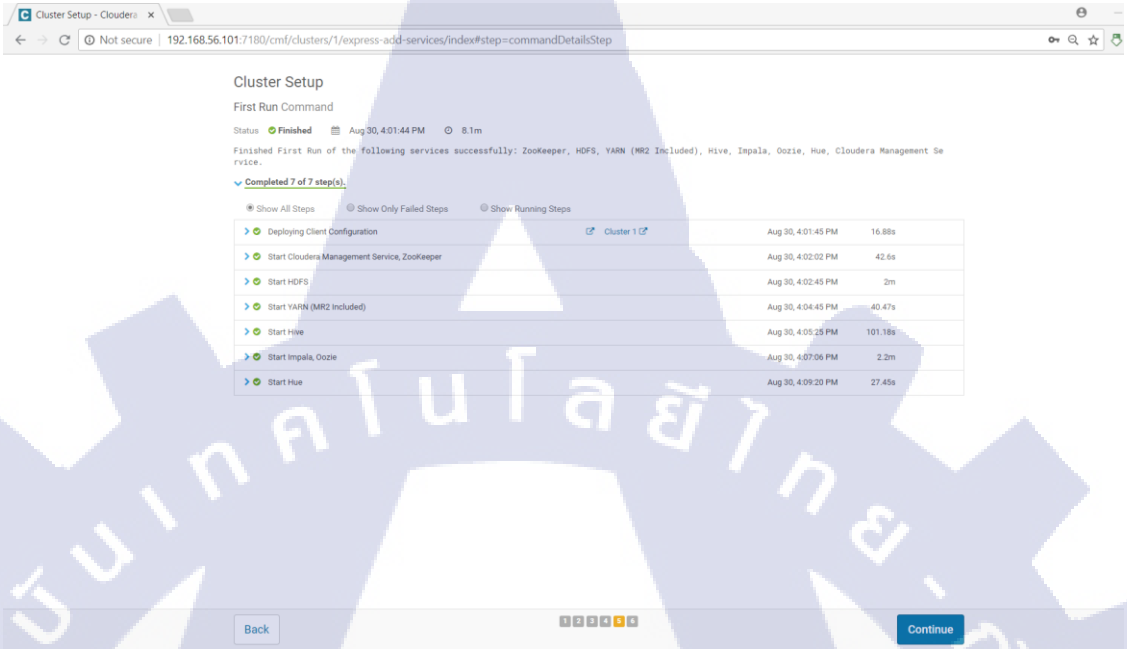
รูปที่ 4.31 หน้า Review Changes

ต่อมาจะเจอหน้า First Run Command ดังรูปที่ 4.32 clouderaจะทำการrun service ที่ผู้จัดทำเลือกมาทั้งหมด



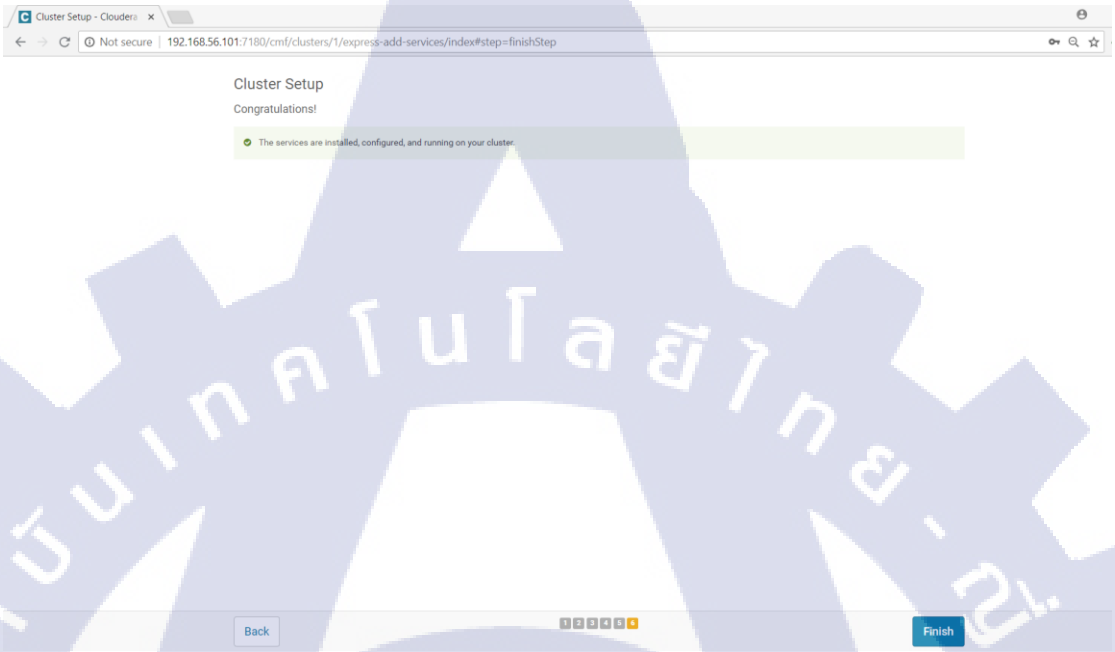
รูปที่ 4.32 หน้า First Run Command

เมื่อเปิด Service ทั้งหมดจะได้ดังรูปที่ 4.33 และกด Continue ไปยังหน้าต่อไป



รูปที่ 4.33 Run service success

ต่อมาเมื่อติดตั้งเสร็จจะเจอหน้า Congratulation ดังรูปที่ 4.34 ให้กด Finish เป็นอันเสร็จสิ้นการติดตั้ง Cloudera และ CDH



รูปที่ 4.34 Service are installed

4.1.2 ติดตั้งโปรแกรม StreamSets ลงบน Cloudera Manager

เป็นขั้นตอนการลงโปรแกรม StreamSets ซึ่งเป็นโปรแกรม open source ใช้ทำ ETL ที่มีความยืดหยุ่นสูง เมื่อต้องการทำETL Big Data นั้นก็ต้องทำให้อยู่ในสภาพแวดล้อมเดียวเพื่อการทำงานมีประสิทธิภาพสูงสุดด้วย Cloudera จึงทำการติดตั้ง StreamSets บน Cloudera ซึ่ง Cloudera มีตัวเชื่อมต่ออยู่แล้ว วิธีการติดตั้ง

Document:https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Installation/CMInstall-Overview.html#concept_nb5_c3m_25 หรือขั้นตอนต่อไปนี้

4.1.2.1 การเตรียมไฟล์ และนำไปเก็บใน Directory /opt/cloudera/parcel-repo

1) StreamSets Parcel

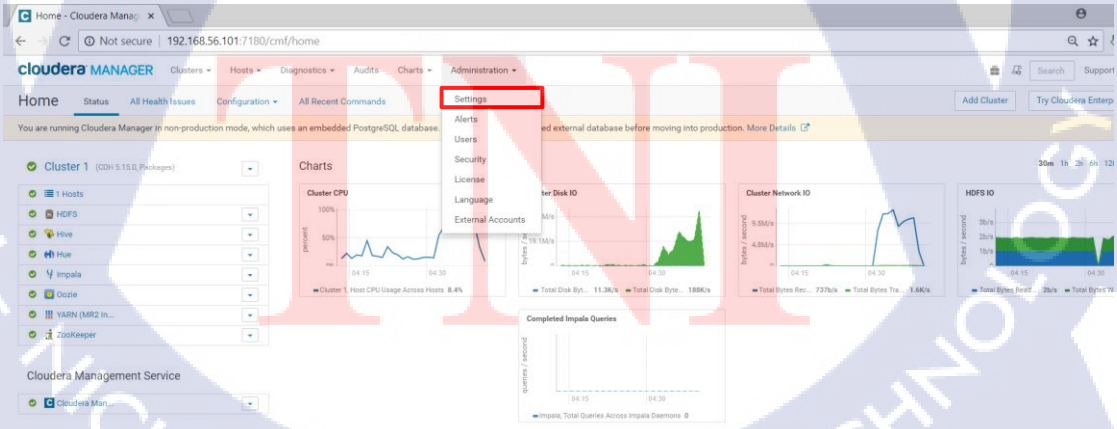
ดาวน์โหลดจาก https://archives.streamsets.com/index.html นำไปใส่ใน Directory /opt/cloudera/parcel-repo มีสองไฟล์ได้แก่ RHEL 7 และ SHA

2) StreamSets CSD

เป็นตัวที่ทำให้ Cloudera เชื่อมต่อกับ StreamSets ดาวน์โหลดจาก https://streamsets.com/opensource/?utm_source=Cloudera_Downloads_Parcel&utm_campaign=Cloudera#smaller_download นำไปใส่ในDirectory /opt/cloudera/csd

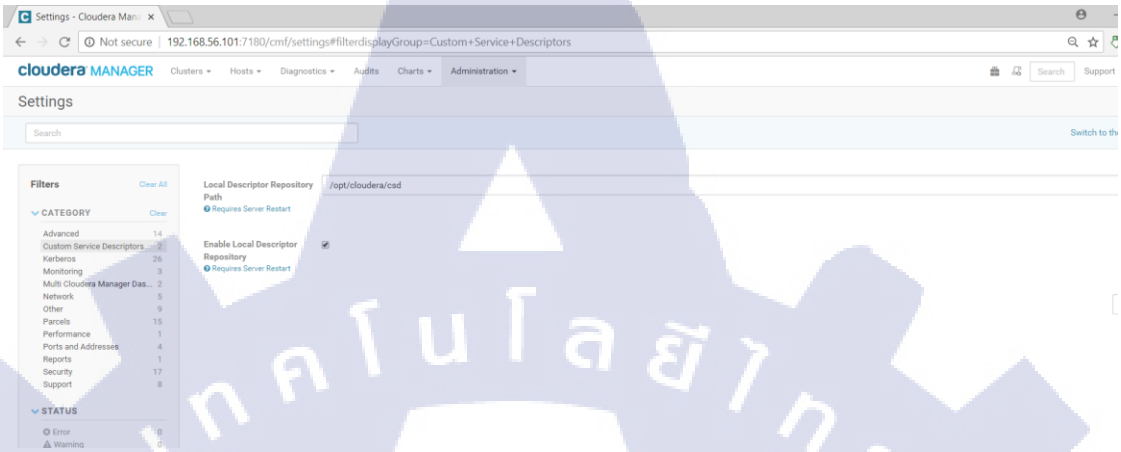
4.1.2.2 ทำการตรวจสอบ path ของ CSD ว่าตรงกับ path ที่นำ CSD ไปใส่ไว้หรือไม่

ไปที่ Cloudera Manager http://localhost:7180 เลือก Administration > Settings ดังรูปที่ 4.35



รูปที่ 4.35 Administration_Settings

ให้เลือก Filter Custom Service Descriptors category ดัง และตรวจสอบช่อง Local Descripter Repository ว่าตรงกันหรือไม่



รูปที่ 4.36 Custom Service Description

4.1.2.3 ตั้งค่าไฟล์ CSD ให้ cloudera เป็น ownership และตั้ง permission

chown cloudera-scm:cloudera-scm /opt/cloudera/csd/STREAMSETS*.jar

chmod 644 /opt/cloudera/csd/STREAMSETS*.jar

และ restart Cloudera Manager Server ถ้าเปิดagent อยู่ให้ปิดก่อนแล้วค่อยrestart

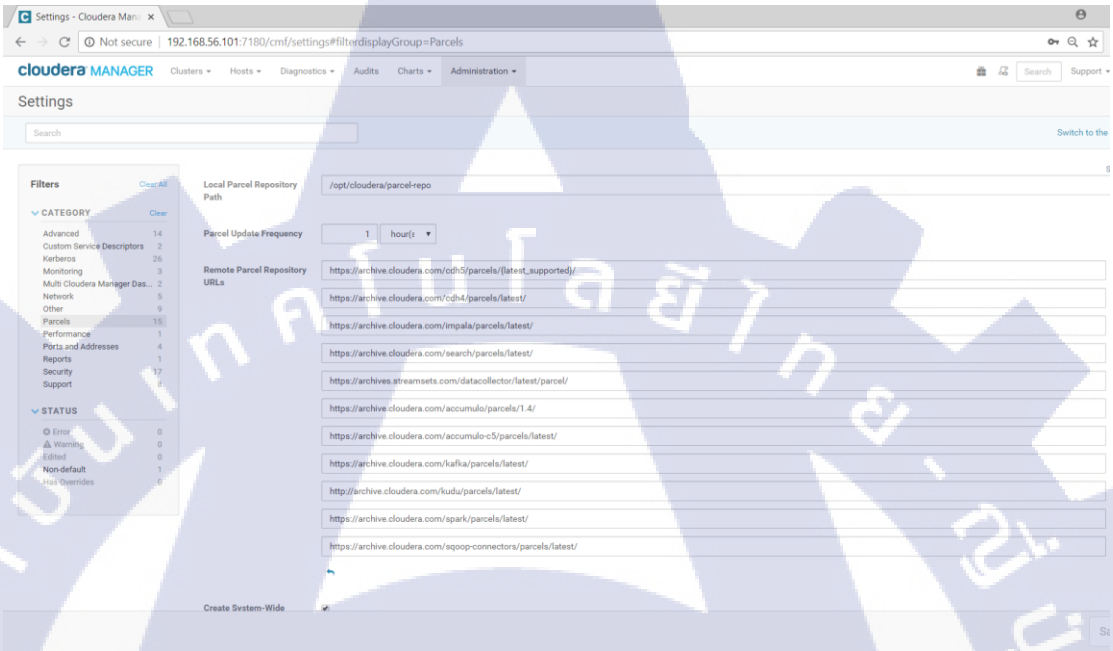
systemctl stop cloudera-scm-agent

systemctl restart cloudera-scm-server

systemctl start cloudera-scm-agent

4.1.2.4 ตรวจสอบ path cloudera parcel

ตรวจสอบว่า Directory /opt/cloudera/parcel-repo ได้ใช่จริงหรือไม่ ไปที่ Cloudera Manager กด Administration > Settings เลือกตรง Filter Parcels category ดังรูปที่ 4.37



รูปที่ 4.37 Path of parcel

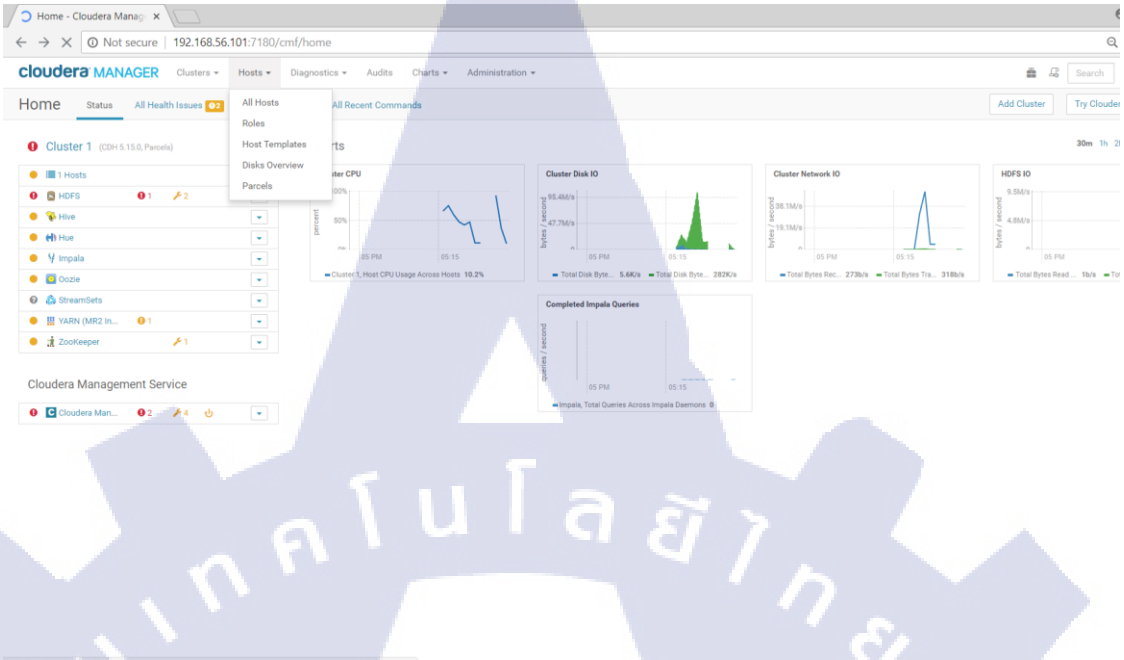
4.1.2.5 ตั้งค่าไฟล์ parcel ให้ cloudera เป็น ownership

chown cloudera-scm:cloudera-scm /opt/cloudera/parcel/Repo

/STREAMSETS_DATACOLLECTOR*

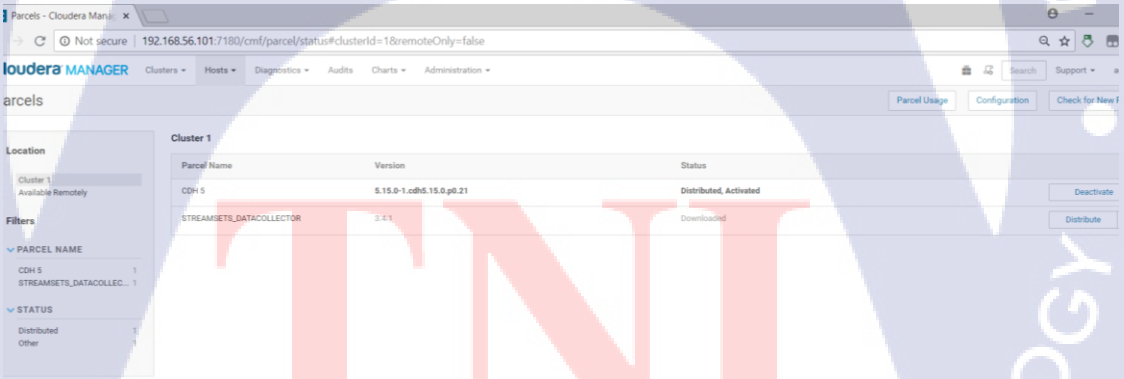
4.1.2.6 Distribute และ Active StreamSets

ไปที่ Cloudera Manager แถบ menu bar เลือก Hosts>Parcels ดังรูปที่ 4.38



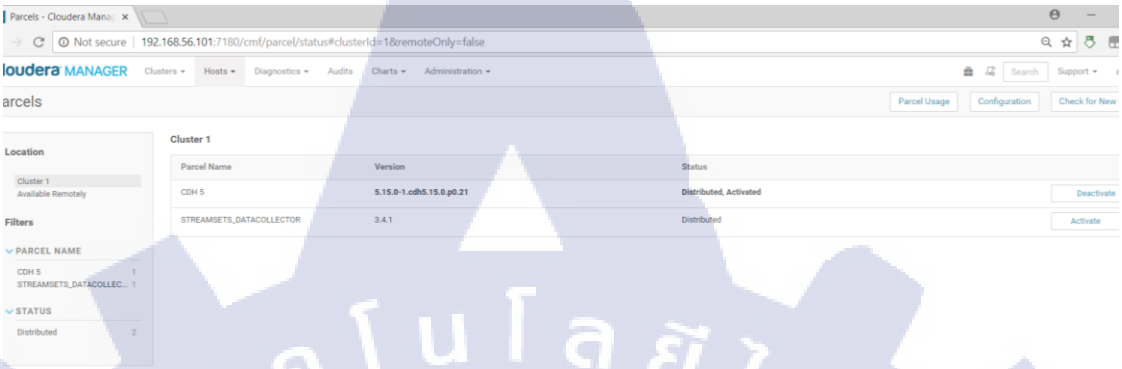
รูปที่ 4.38 Host_Parcels

จะมี STREAMSETS_DATACOLLECTOR ให้เลือก Distribute ดังรูปที่ 4.39



รูปที่ 4.39 Parcel_distribute

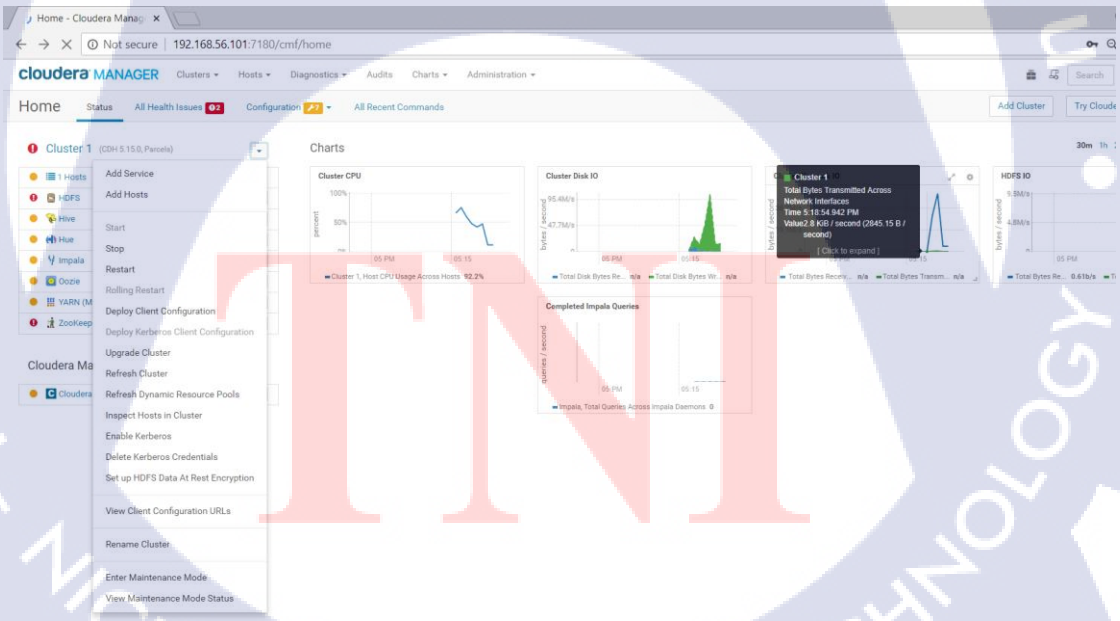
ให้กด Distribute จะทำการDistribute parcel เมื่อเสร็จแล้ว ให้กด Activate ดังรูปที่ 4.40 เมื่อกดจะมีpop up ถาม Are you sure? ให้เลือก OK เมื่อactivate เสร็จจะสามารถกด Deactivate ได้



รูปที่ 4.40 Parcels_activate

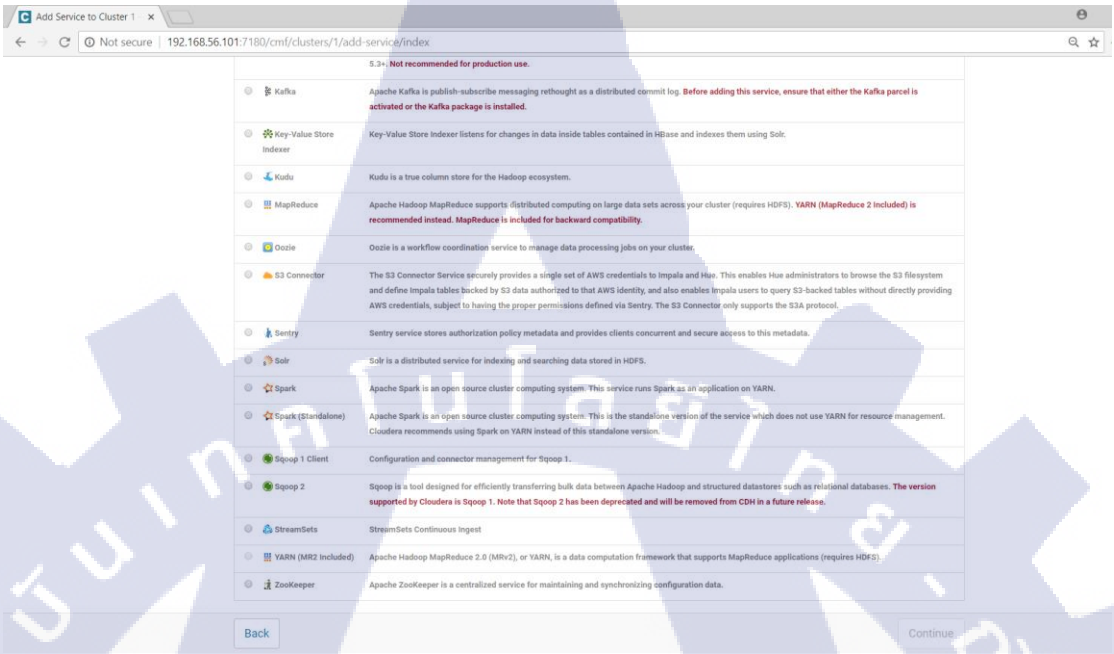
4.1.2.7 เพิ่ม service ให้ cloudera Manager

กดปุ่ม menu drop down list > Add Service ดังรูปที่ 4.41



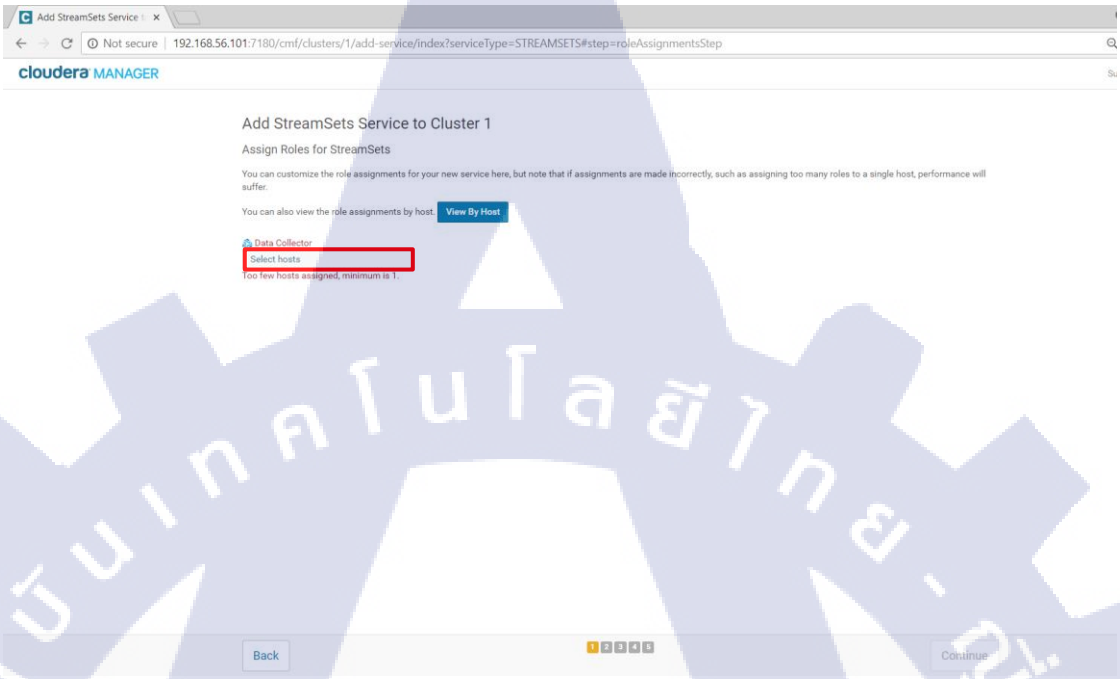
รูปที่ 4.41 Add service

ให้เลือก Service StreamSets แล้วกด Continue ดังรูปที่ 4.42



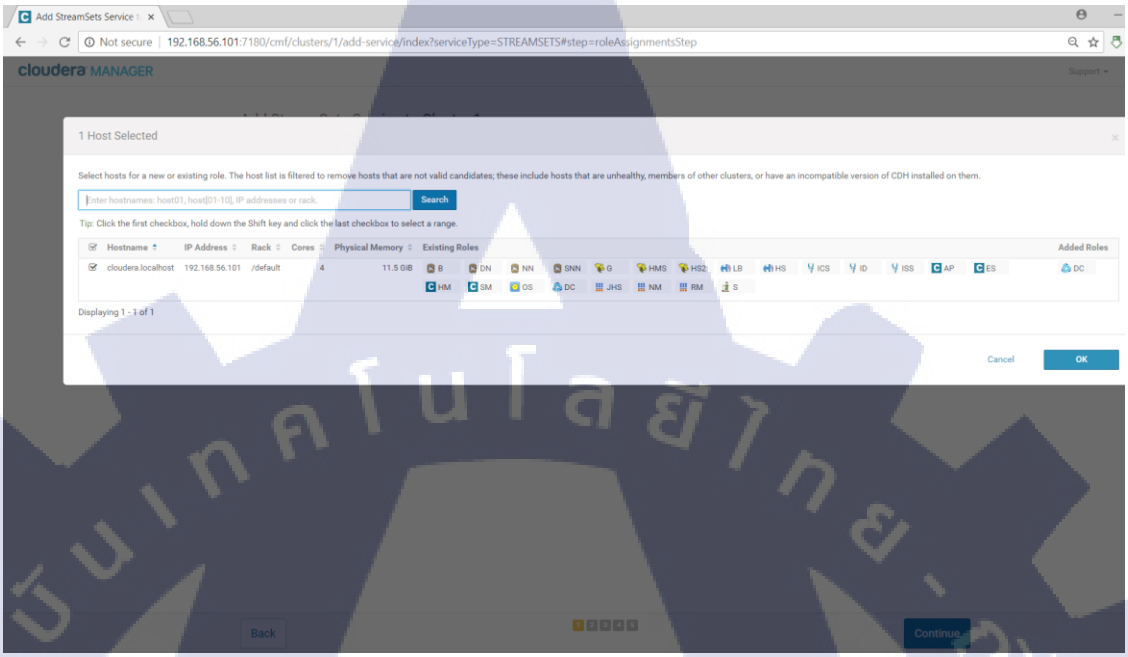
รูปที่ 4.42 เลือก Service StreamSets

ต่อจะมาหน้าให้เลือก hosts ให้กดที่ Select hosts ดังรูปที่ 4.43



รูปที่ 4.43 เลือก Assign role for StreamSets

เมื่อกดแล้วให้เลือก hosts ที่ต้องการดังรูปที่ 4.44 เสร็จให้กด OK และ กด Continue

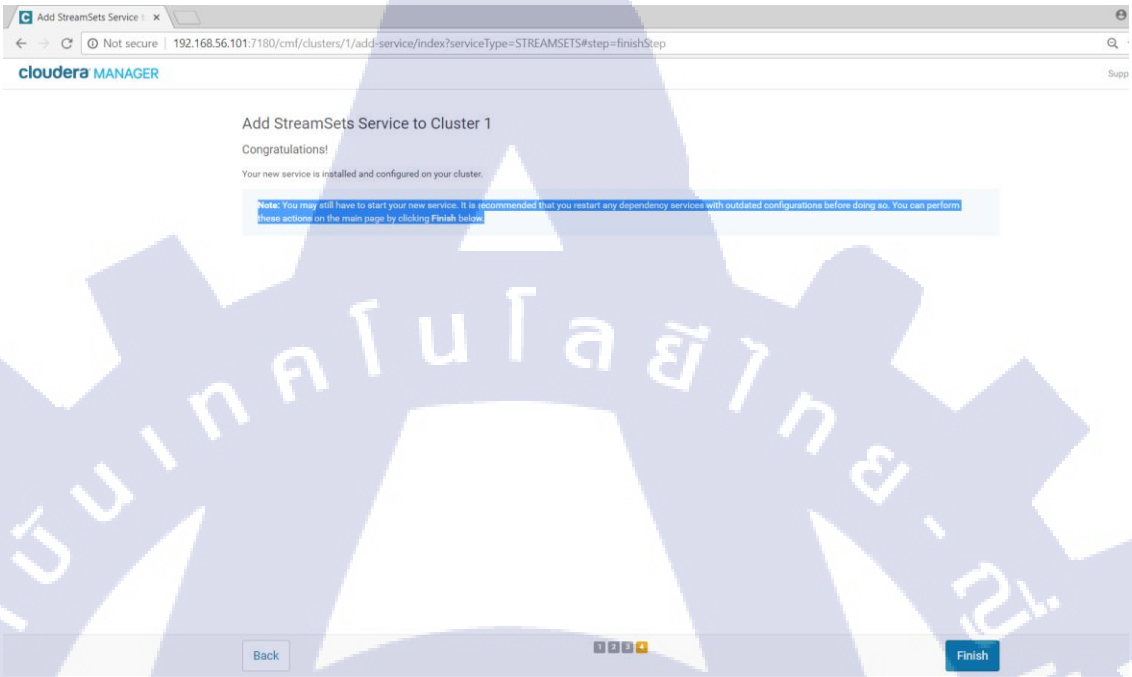


รูปที่ 4.44 Select hosts

Cloudera manager จะทำการ run command เพื่อเปิด service เมื่อเสร็จกด continue ดังรูปที่ 4.45

รูปที่ 4.45 Cloudera start service StreamSets

ต่อมาจะเป็นหน้า Congratulation ถือเป็นการเสร็จสิ้นการติดตั้งStreamSets บน Cloudera ดังรูปที่ 4.46



รูปที่ 4.46 ติดตั้ง StreamSets สำเร็จ

4.1.3 การนำข้อมูลเข้า Hadoop (Hadoop Distrubute File System)โดยการใช้ Hue

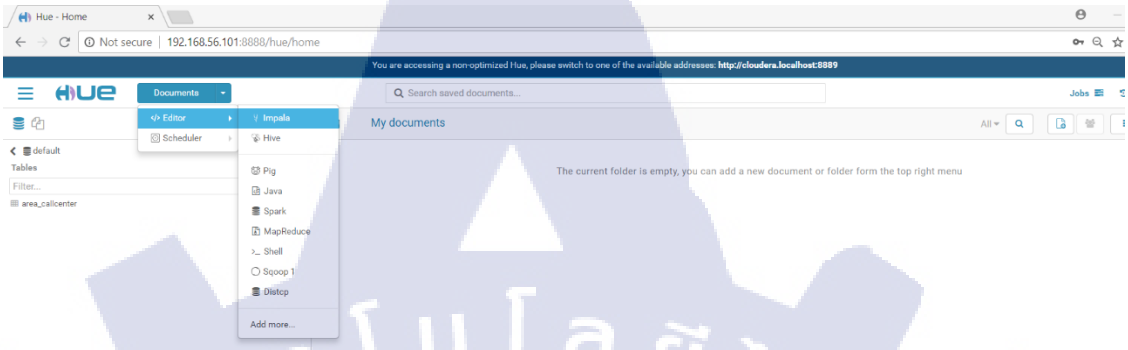
เป็นขั้นตอนการนำข้อมูลเข้าที่เก็บข้อมูลของ Hadoop คือ Hadoop Distribute File System โดยใช้Hue (Hadoop User Exprience) และ command line Linux

4.1.3.1 เข้าไปยังหน้าHue

http://localhost:8888 ผู้จัดทำเข้าโดยใช้ ip คือ http://192.168.56.101 ดังรูปที่ 4.47 เมื่อเข้าครั้งแรกจะให้ทำการตั้งsuperuser ทันที

รูปที่ 4.47 Hue login

ต่อมาจะเป็นหน้า Home ของ Hue สามารถเลือก Document เพื่อดู service ที่มีอยู่ดังรูปที่ 4.48

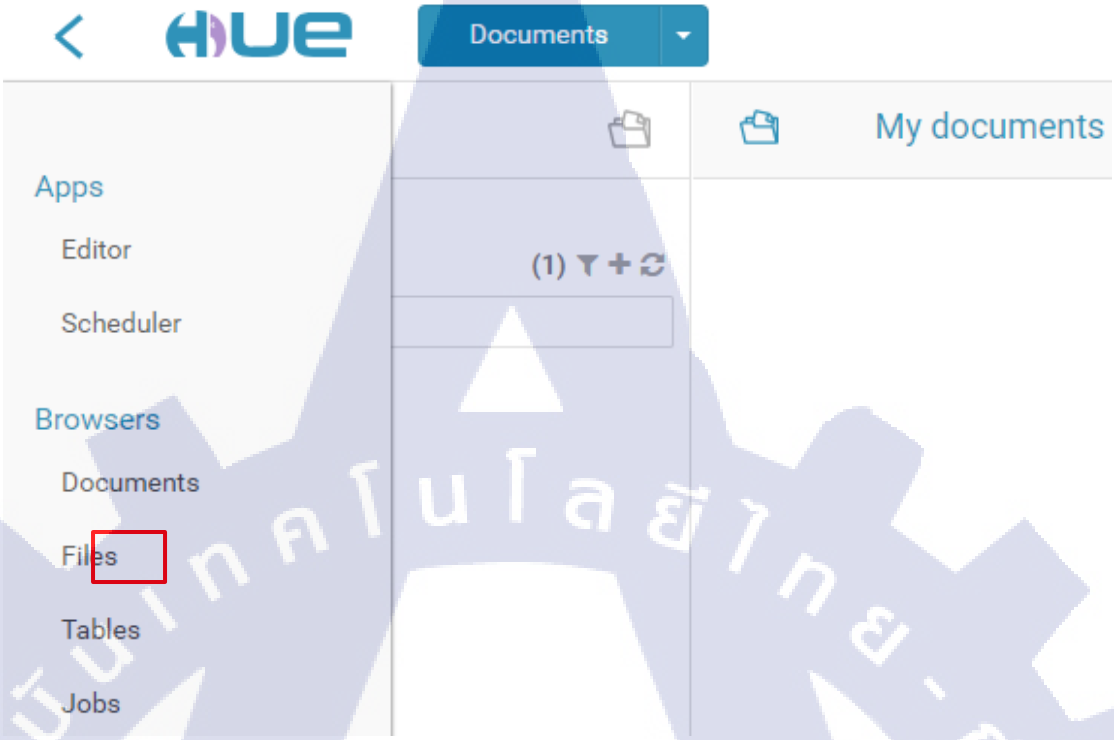


รูปที่ 4.48 Home_Hue

4.1.3.2 เข้าไปยังหน้า HDFS

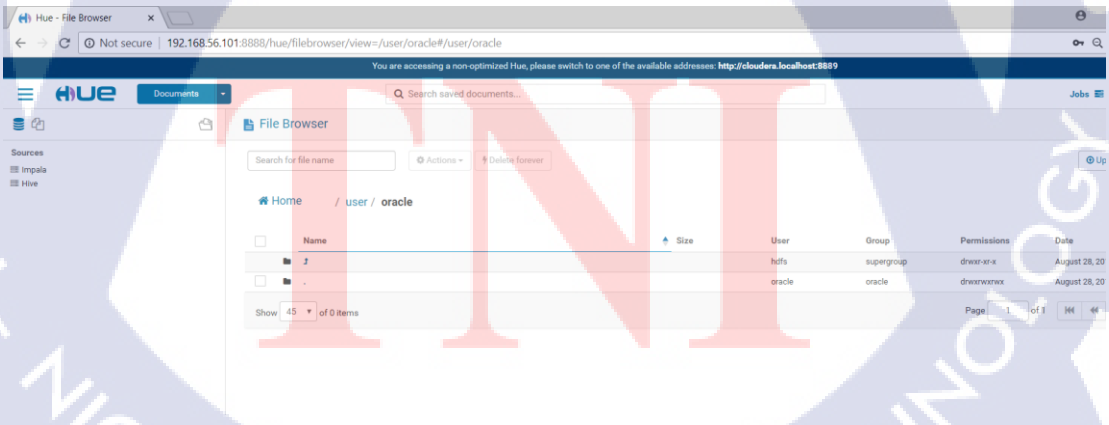
ต่อมาเลือกปุ่มไอคอน 3 ขีดด้านซ้ายมือดังรูปที่ 4.49 เมื่อเลือกจะมีเมนูต่างๆ ให้เลือก Files ดังรูปที่ 4.50

รูปที่ 4.49 ไอคอนเมนู 3 ขีดของ hue

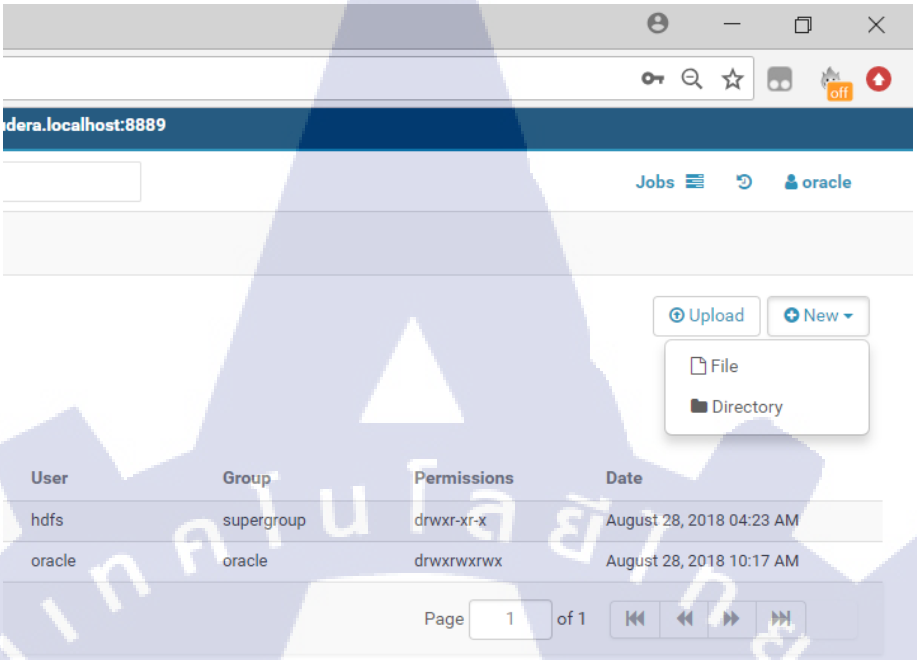


รูปที่ 4.50 Menu_File

เมื่อเข้ามาจะมาอยู่หน้า Home ของ HDFS คือ /user/oracle ดังรูปที่ 4.51 สามารถกดupload หรือ กด new เพื่อสร้าง Directory หรือสร้างไฟล์ได้ดังรูปที่ 4.52



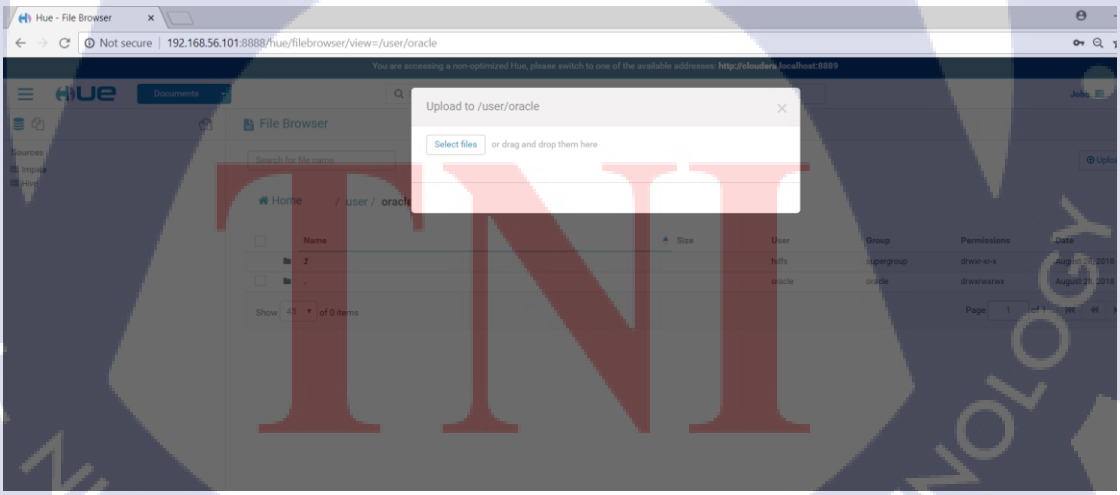
รูปที่ 4.51 หน้าเก็บ File hdfs



รูปที่ 4.52 สร้าง File หรือ Directory

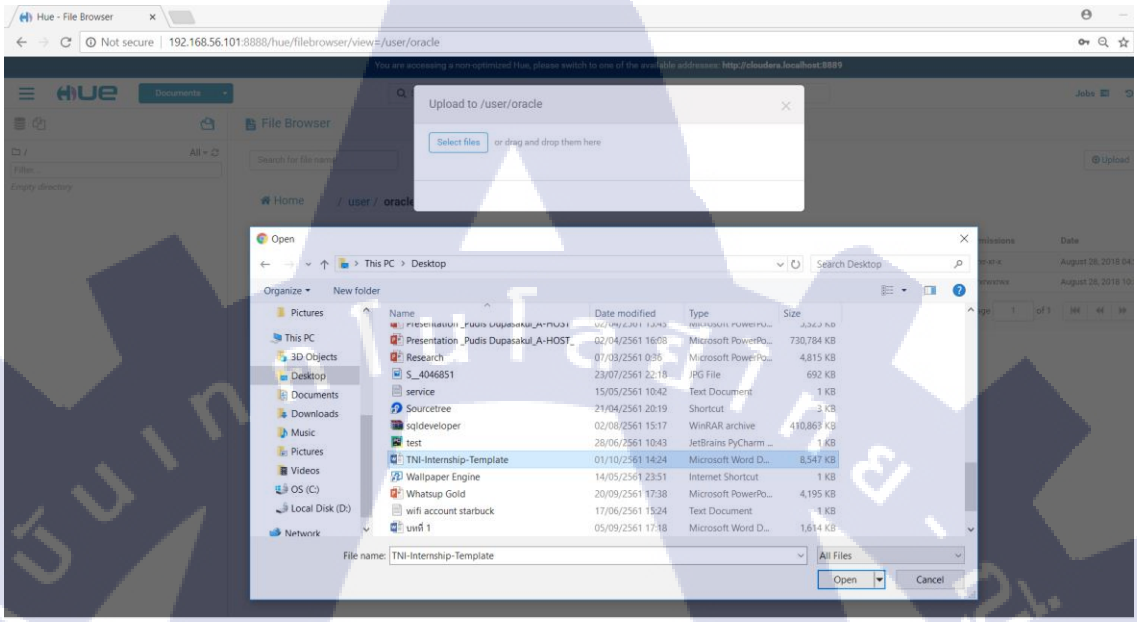
4.1.3.3 upload file เข้า HDFS

เมื่อกด upload จะขึ้น popup ให้เราเลือกไฟล์ดังรูปที่ 4.53 สามารถเลือกไฟล์ตามที่ต้องการ



รูปที่ 4.53 หน้า popup select files

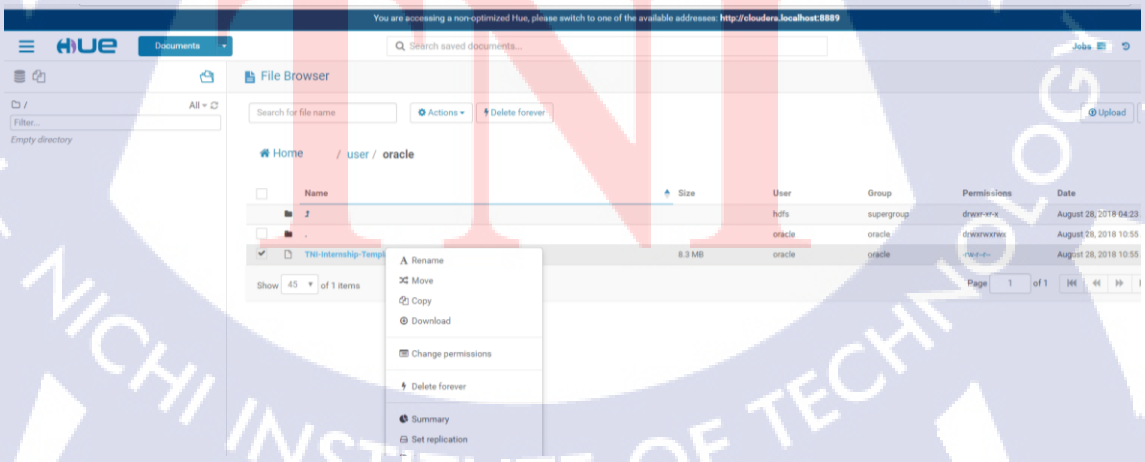
เมื่อกด Select Files จะขึ้นหน้าให้เราเลือกไฟล์ เมื่อเลือกเสร็จกด open เพื่อ upload ดังรูปที่ 4.54



รูปที่ 4.54 หน้าเลือก File เพื่ออัพโหลด

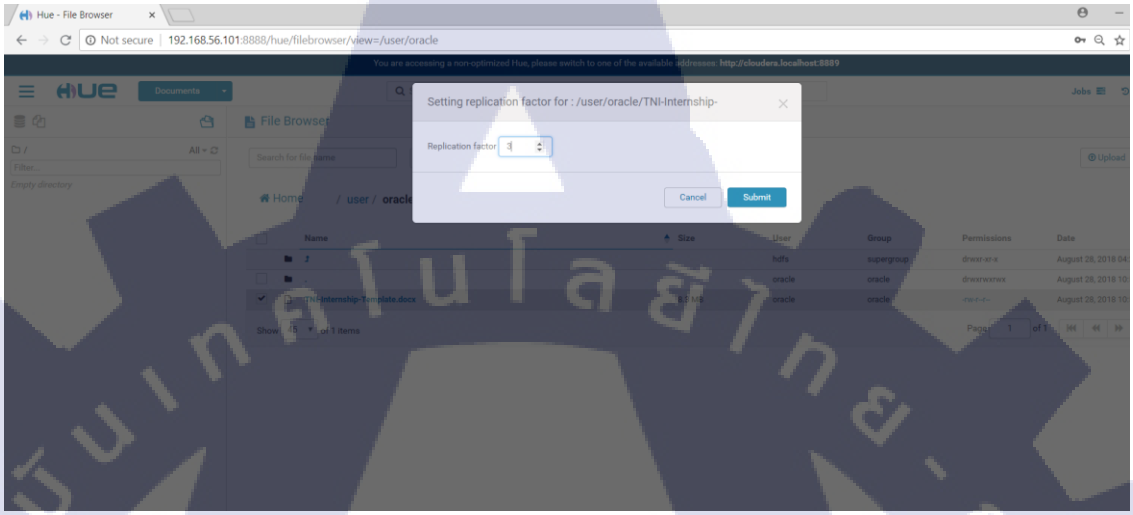
4.1.3.4 ตั้งค่า Files

สามารถตั้งค่าการ replication,permission,compress และอื่นได้ด้วยการคลิกขวาดังรูปที่ 4.55



รูปที่ 4.55 คลิกขวาเพื่อตั้งค่าไฟล์

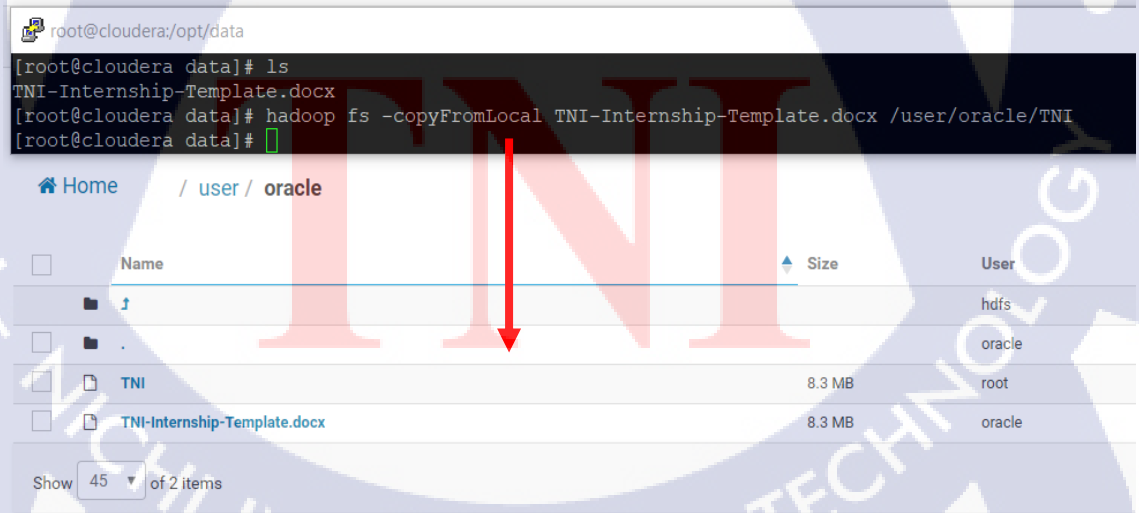
เลือก Set replication แล้วจะขึ้นดัง ค่าพื้นฐาน replication คือ 3 โดยพื้นฐานของHadoop จะต้องมีName node 1 เครื่อง และมี Data Node 3 เครื่อง โดยการ replication 3 นั้นจะพอดีกับ Data node



รูปที่ 4.56 เลือก replication

หรือสามารถนำข้อมูลเข้าโดยใช้ command Linux

#hadoop fs -copyFromLocal [/path/file] [/path/hdfs]



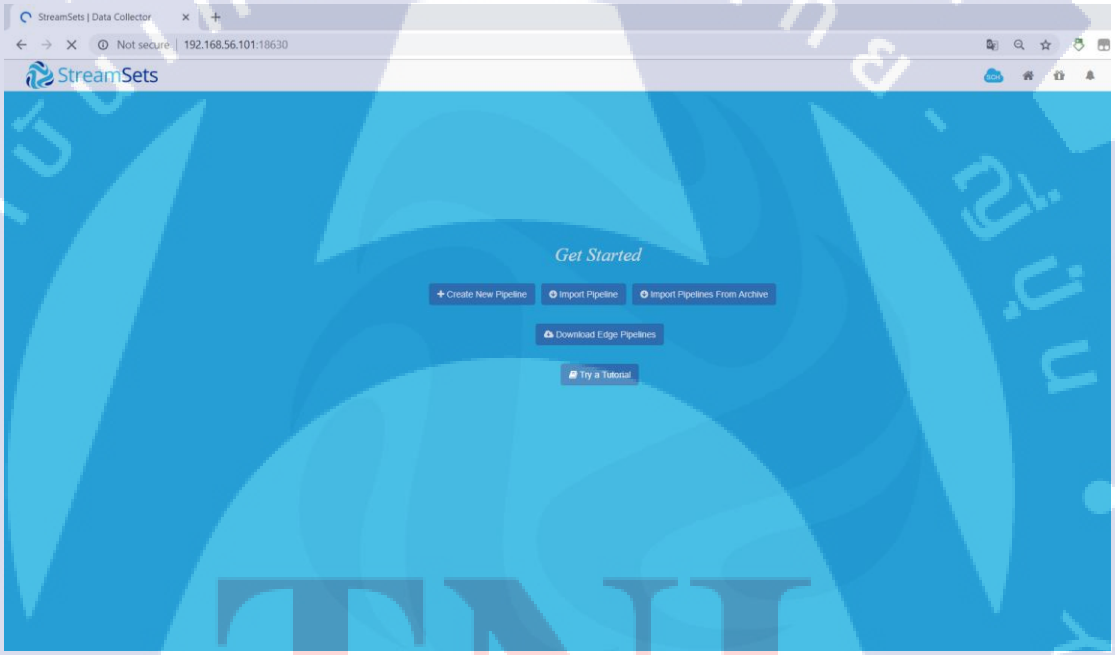
รูปที่ 4.57 ตัวอย่างการนำข้อมูลเข้าด้วย command line

4.1.4 การทำ ETL ด้วย StreamSets

เป็นขั้นตอนการทำETL คือการนำข้อมูลจากต้นทาง ไปปรับปรุงและส่งไปยังที่เก็บข้อมูลปลายทางในที่นี้เก็บใส่ใน Hadoop Distribute File System การทำงานของ StreamSets จะเป็น 3 ส่วนหลักๆ คือ 1.Origin คือต้นทางของข้อมูล 2.Processors คือส่วนประมวลผล ปรับปรุงข้อมูล 3.Destination คือส่วนการส่งข้อมูลไปยังปลายทาง ในที่นี้จะทำการสร้าง flow การทำงานพื้นฐานที่ง่ายต่อการเข้าใจและครบกระบวนการการทำ ETL โดยจะใช้ File CSV ในการทำ

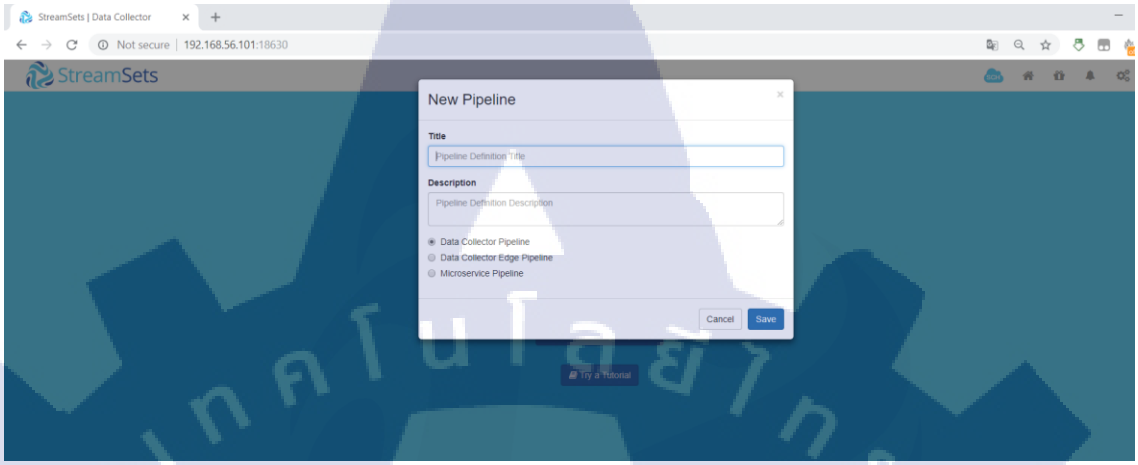
4.1.4.1 ใช้ Browser เข้าไปยัง StreamSets

http://localhost:18630 ในที่นี้ผู้จัดทำใช้ http://192.168.56.101 เมื่อเข้าครั้งแรกจะเจอหน้าดังรูปที่ 4.58



รูปที่ 4.58 เข้าStreamSets ครั้งแรก

มีแบบฝึกหัดให้ทำอยู่ด้านล่างคือปุ่ม Try a Tutorial ในที่นี้เราจะเริ่มทำการสร้าง pipe line เพื่อเริ่มการ ETL โดยการกด Create New Pipeline แล้วจะขึ้นดังรูปที่ 4.59



รูปที่ 4.59 สร้าง pipeline ใหม่ StreamSets

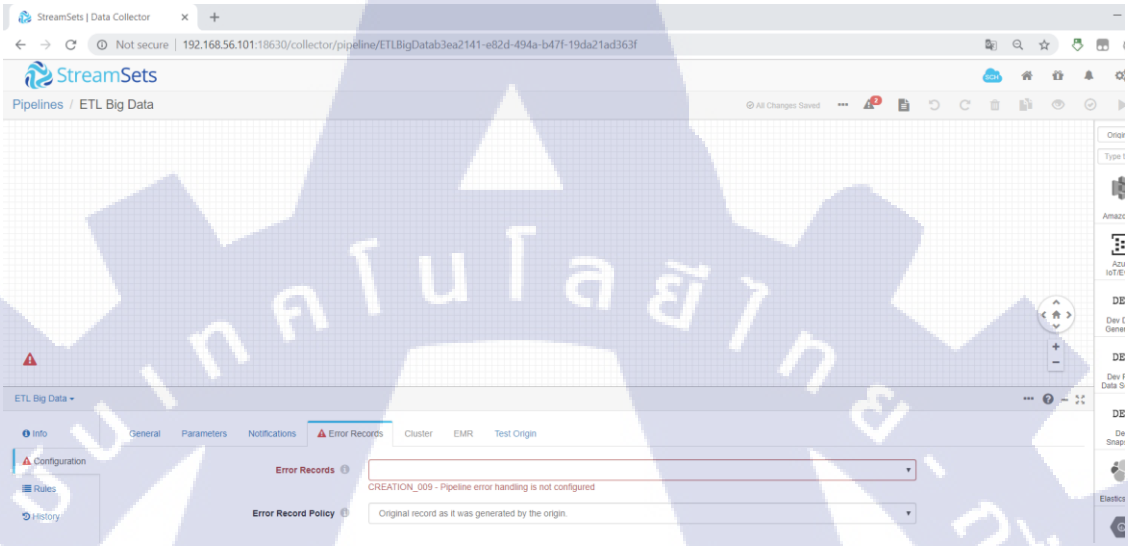
เมื่อใส่ข้อมูลTitle และ Description ก็เลือก Data Collector Pipeline และกด Save จะแสดงหน้าดังรูปที่ 4.60 ส่วนที่ 1 สำหรับว่าง flow การทำงานโดยสามารถนำส่วนที่ 3 ที่เป็น ส่วนของ Stage Library ต่างๆ ลากมาว่างในส่วนที่ 1 ได้ ส่วนที่ 2 คือที่ตั้งค่าของ pipeline และ library ต่าง ๆ

รูปที่ 4.60 Pipeline StreamSets

4.1.4.2 การทำ Flow ETL

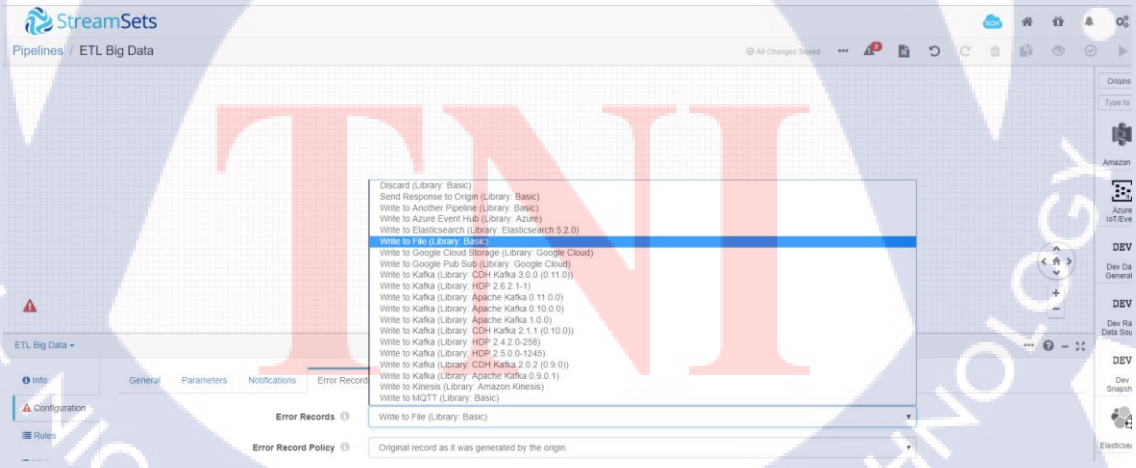
1) ตั้งค่า pipeline

เมื่อCreate new pipeline มาถึงจะเห็นError เครื่องหมายตกใจสีแดง จะเป็นตัวแจ้งเตือนให้เราตั้งค่าให้ถูกต้อง ให้เรากดเลือก Error Records ดัง รูปที่ 4.61



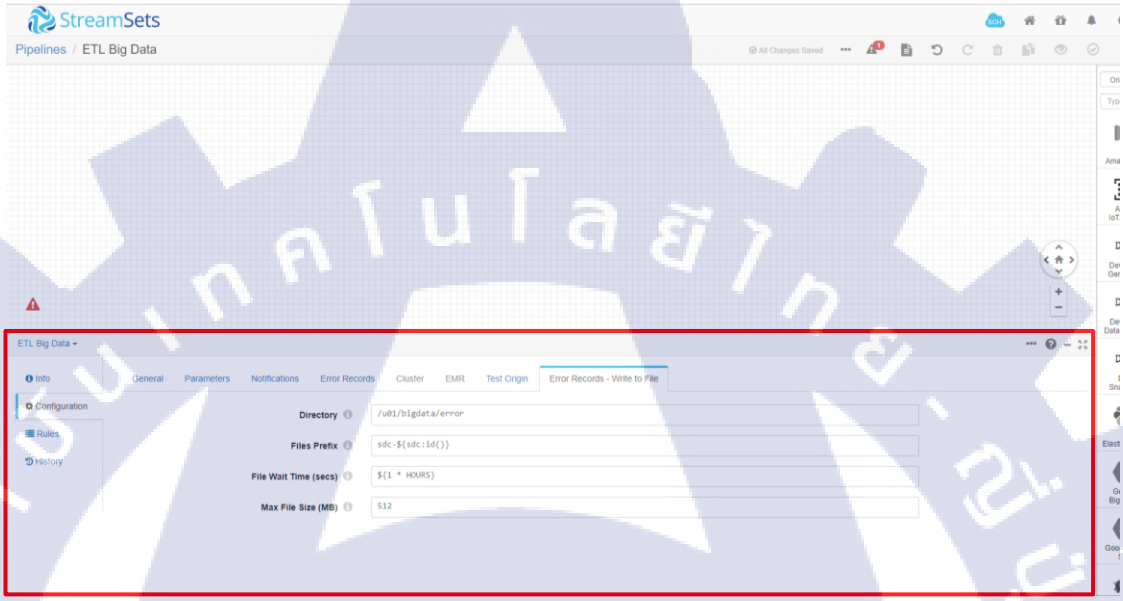
รูปที่ 4.61 เลือก Error Records

เลือก Error Records เป็น Write to File (Library: Basic) ดังรูปที่ 4.62



รูปที่ 4.62 Write to File (Library: Basic)

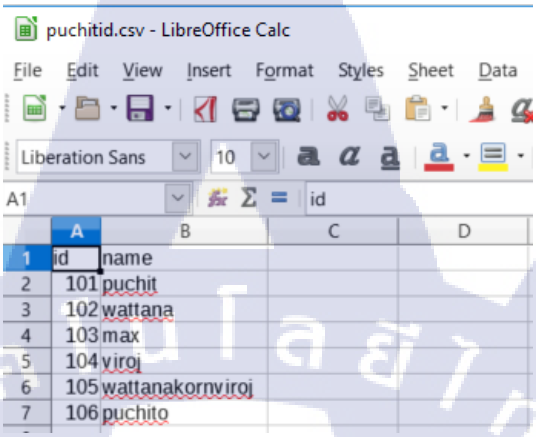
เมื่อกดเลือกเสร็จจะเพิ่มแถบ Error Records – write to Fileขึ้นมาให้เราเลือกและตั้งค่า Directory ที่ต้องการเก็บข้อมูลเมื่อเกิด Error ดัง ถ้าไม่ตั้งค่าจะขึ้นแจ้งเตือน VALIDATION_0007 - Configuration value is required ใต้ช่อง Directory และ Directory ที่เลือกต้องตั้ง permission ให้ write ได้



รูปที่ 4.63 ตั้งค่าที่อยู่เมื่อเขียน File เสร็จ

2) Origin

จะเป็นส่วนต้นทางข้อมูลที่เข้ามาในส่วนนี้มี File CSV โดยข้อมูลที่มีดังรูปที่ 4.64



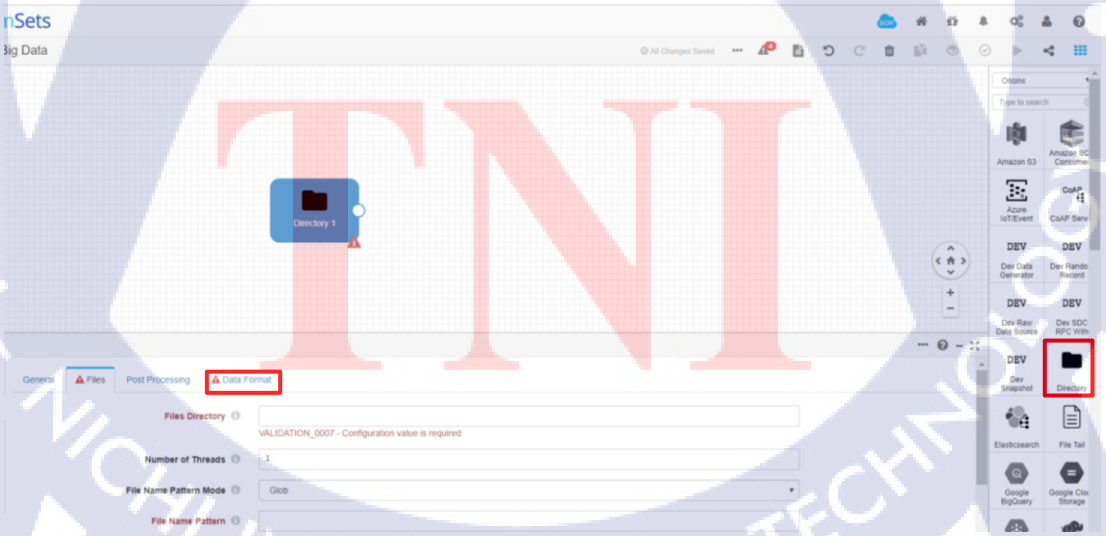
รูปที่ 4.64 ข้อมูลทดสอบในFile CSV

ให้ทำการเลือก Directory ในกรอบแดงกดคลิกซ้ายที่ icon เพียงครั้งเดียวจะได้ดังรูปที่ 4.65 เมื่อเลือกแล้วให้ทำการตั้งค่า path ที่อยู่ของ file และประเภทของไฟล์ผู้จัดทำใส่ดังนี้

Files Directory : /u01/bigdata/data

File name Pattern : *.csv (เลือกทุกไฟล์ที่มี .csv)

และเลือกหัวข้อ Data Format



รูปที่ 4.65 เลือก Origin directory

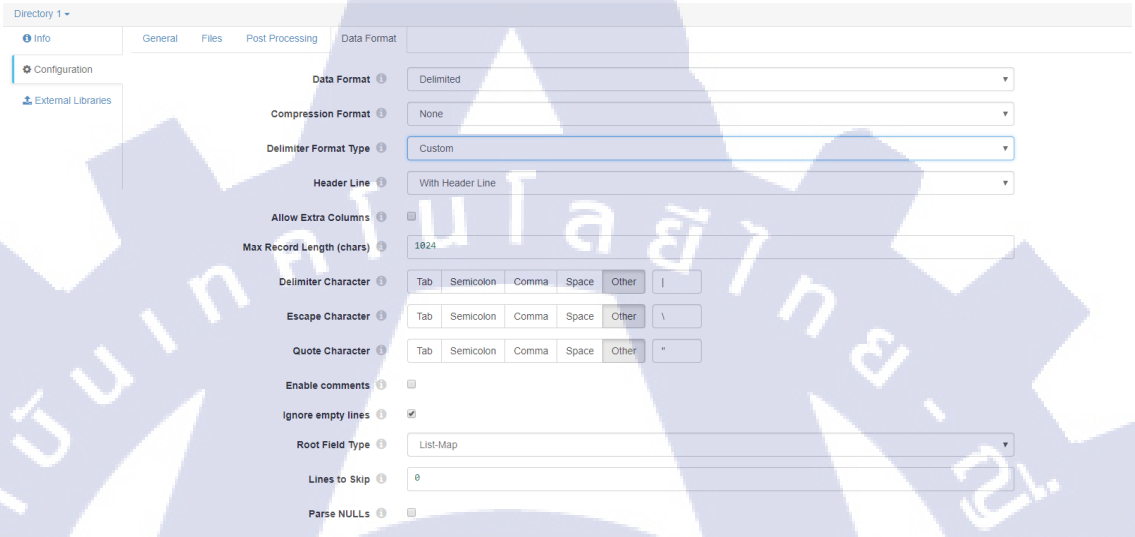
Data Format เลือก Delimited เมื่อเลือกแล้วจะขึ้นมา และตั้งค่าตามนี้

Delimiter Format Type : Custom

Header Line : With Header Line

Delimiter Character : other = | (CSV มี Delimiter Character คือ | (pipeline))

เหมือนดังรูปที่ 4.66

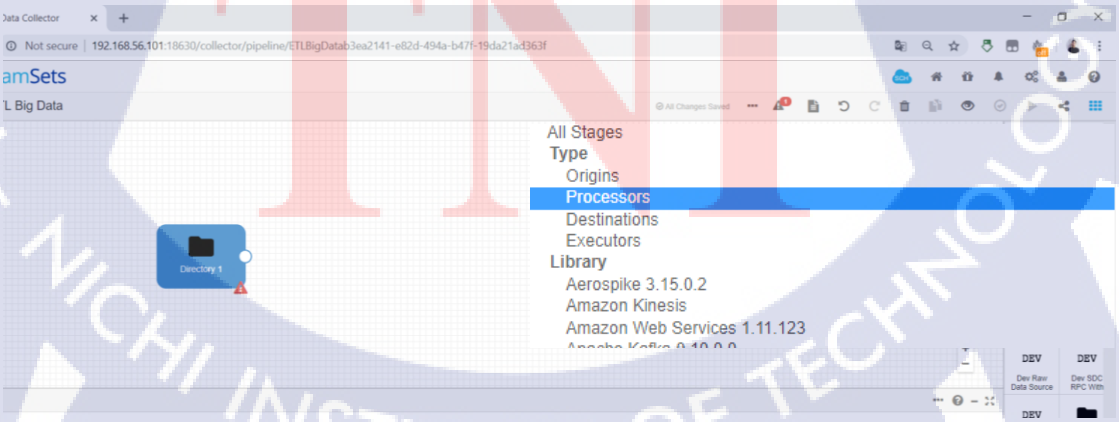


รูปที่ 4.66 Data format Delimited

3 Processors

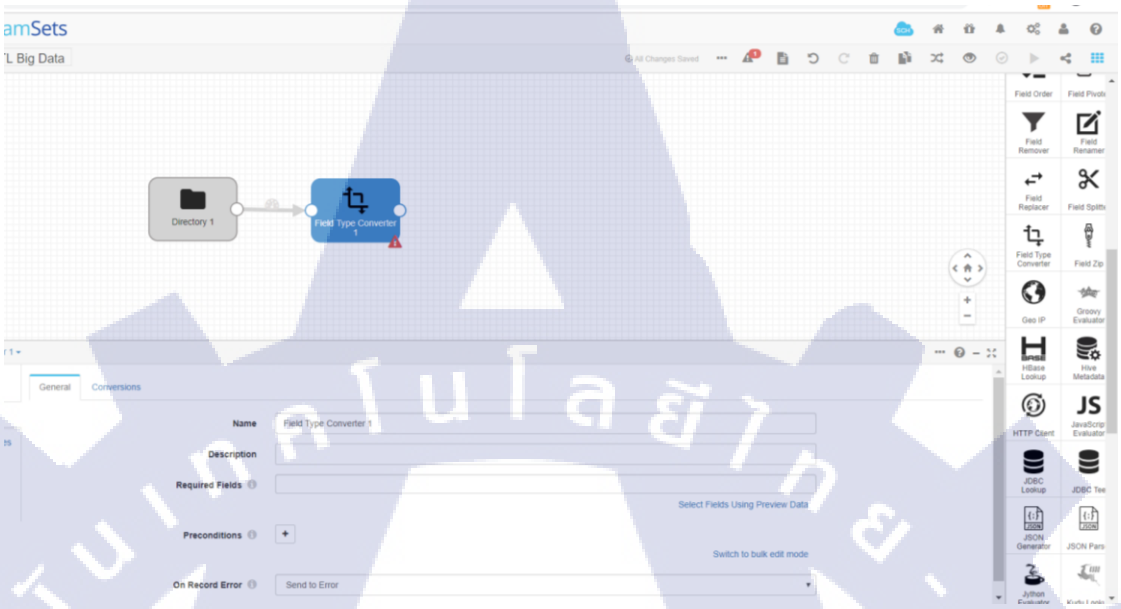
Processors ทำหน้าที่เปลี่ยนแปลงปรับปรุงข้อมูล ให้ทำการเลือก stage เป็น Processorsดัง

รูปที่ 4.67 Processors



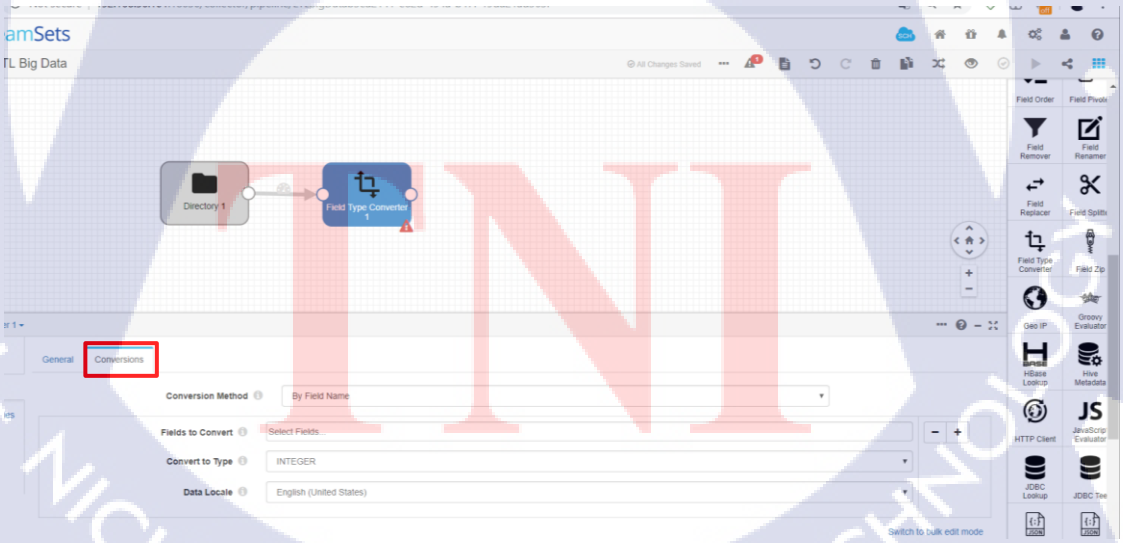
รูปที่ 4.67 Processors

เลือก Field Type Converter จะได้ดังรูปที่ 4.68



รูปที่ 4.68 Field Type Converter

เลือกหัวข้อ Conversions ดังรูปที่ 4.69



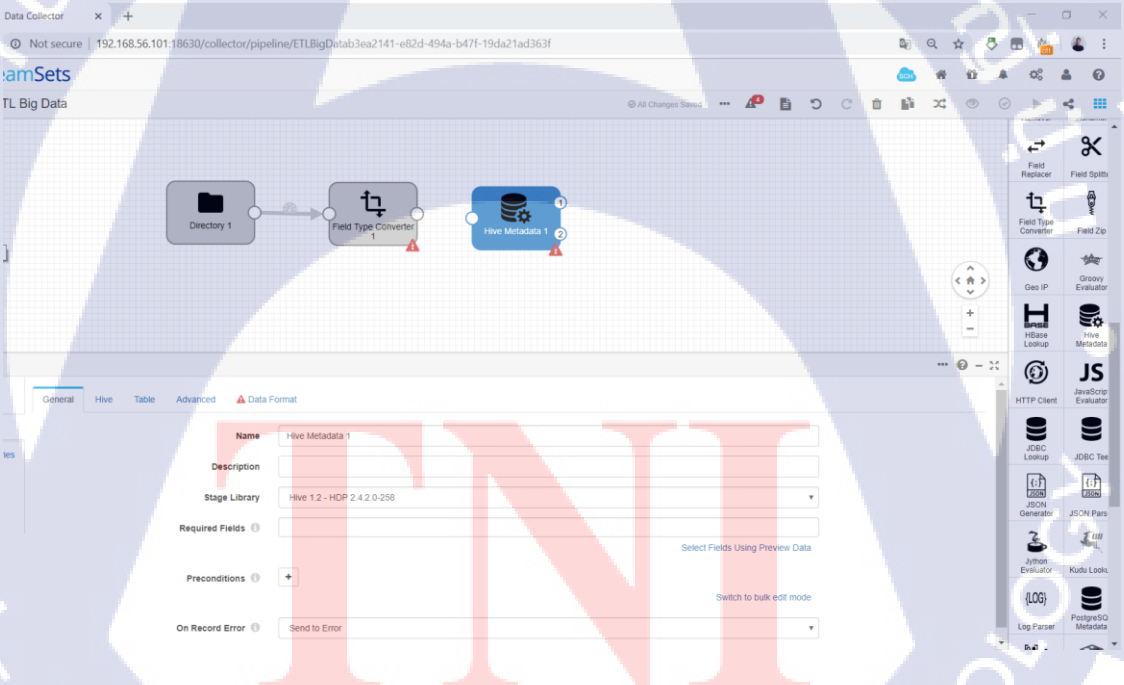
รูปที่ 4.69 Conversions

เลือก Fields to Convert ดังรูปที่ 4.70 และเลือก /id หรือพิมพ์ก็ได้

Convert to Type : INTEGER

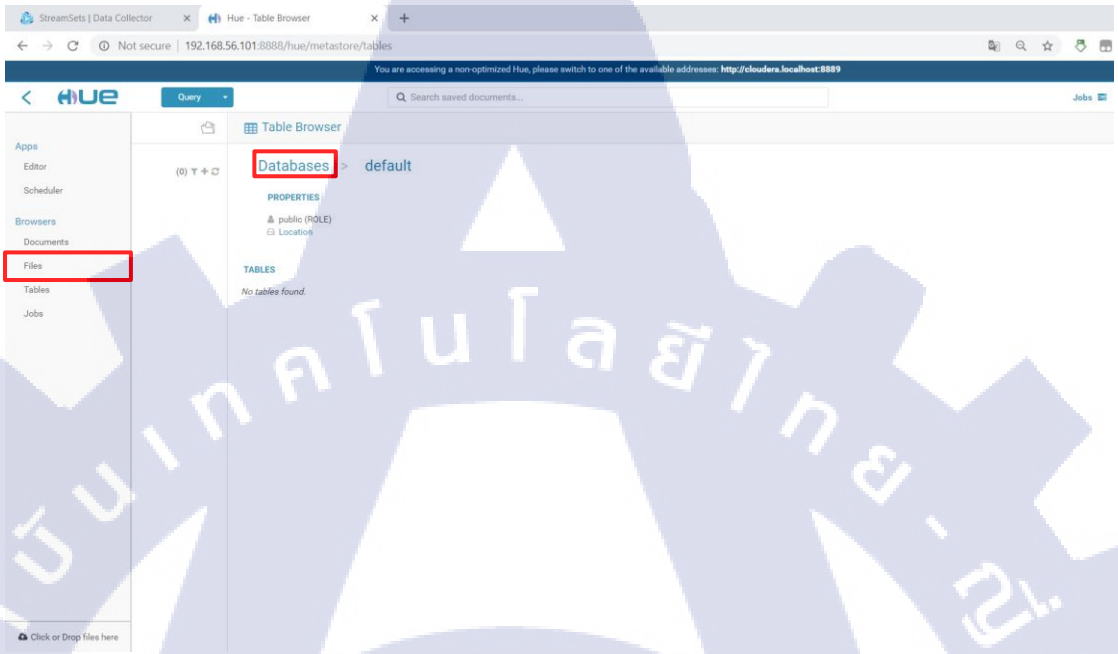
รูปที่ 4.70 เลือก Field to Convert

และเลือก processor อีกตัวคือ Hive Metadata จะได้ดังรูปที่ 4.71



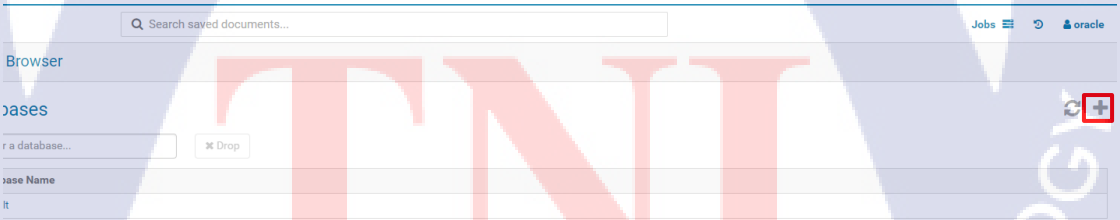
รูปที่ 4.71 Hive Metadata

ก่อนที่จะตั้งค่าเราต้องไปสร้าง table ให้กับ file ก่อนไปที่ hue เลือกเมนูไอคอนสามขีดและเลือก Tables จะได้ดังรูปที่ 4.72 จะมายังหน้า Table Browserและกดที่ Databases



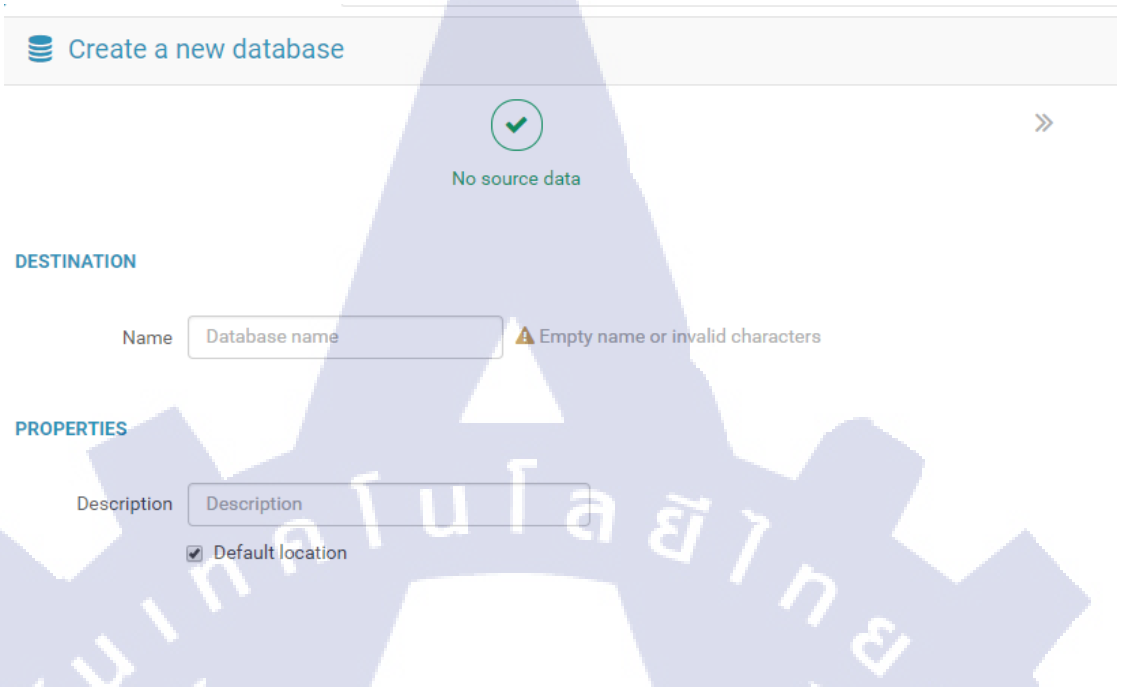
รูปที่ 4.72 Hue Tables

เมื่อเข้ามาแล้วจะดังรูปที่ 4.73 และกดเครื่องหมายบวกเพื่อสร้าง Databases



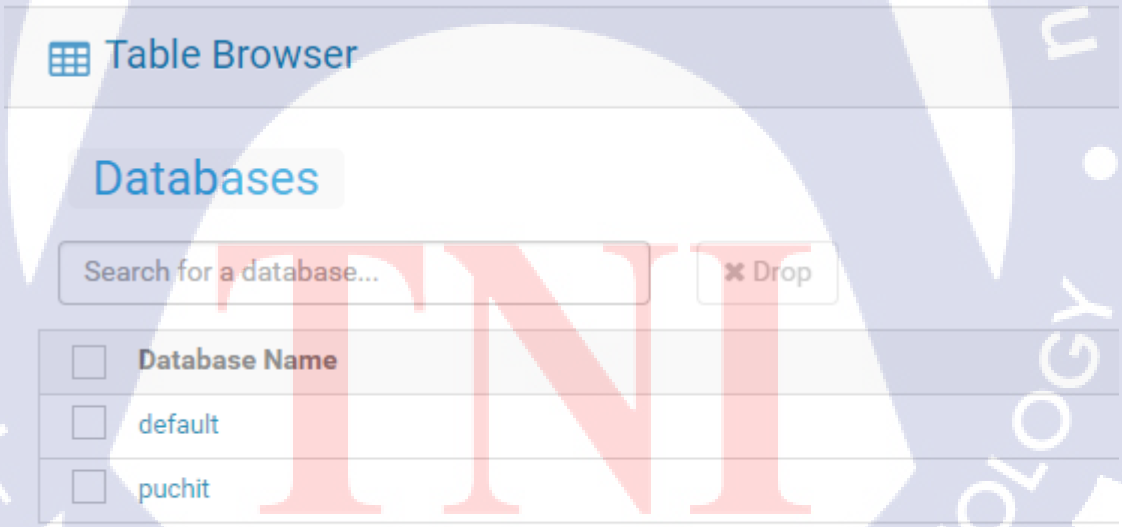
รูปที่ 4.73 Table Databases

เมื่อเลือกแล้วจะมายังหน้า Create a new database ดังรูปที่ 4.74 ให้ใส่ชื่อDatabase ส่วน Description จะใส่ไม่ใส่ก็ได้ แล้วกด Submit



รูปที่ 4.74 Create a new database

เมื่อสร้างเสร็จแล้วจะขึ้นดังรูปที่ 4.75



รูปที่ 4.75 สร้าง database เรียบร้อย

กลับไปยังStreamSets stage Hive Metadata เลือกหัวข้อ General Stage Library

ให้เลือก CDH 5.xx.xx

เลือกหัวข้อ Hive

JDBC URL : jdbc:hive2://<host>:<port>/default

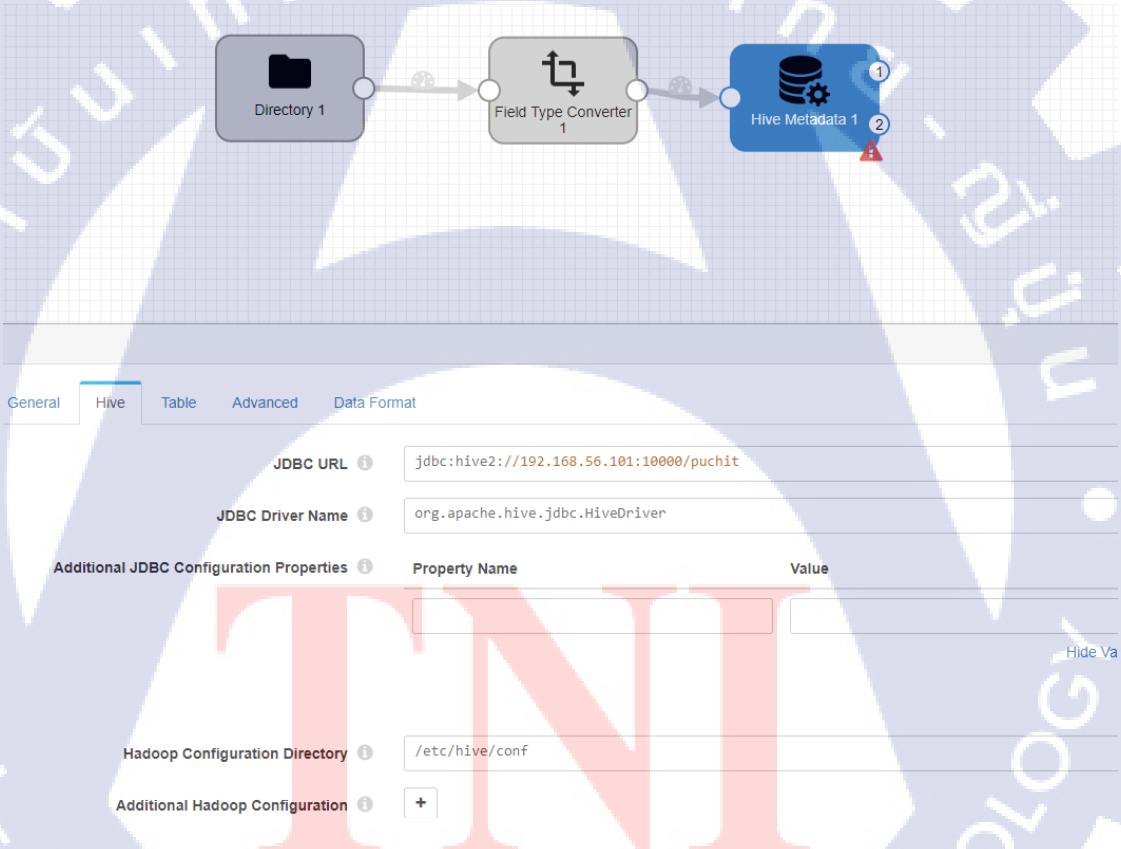
ผู้จัดทำใส่ JDBC URL: jdbc:hive2://192.168.56.101:10000/puchit

เลือกหัวข้อ Table

Database Expression: puchit (คือชื่อdatabase ที่ผู้จัดทำสร้าง)

Table Name : puchitid (ตั้งชื่อ table ตามที่ต้องการ hive metadata จะทำการสร้าง table ให้)

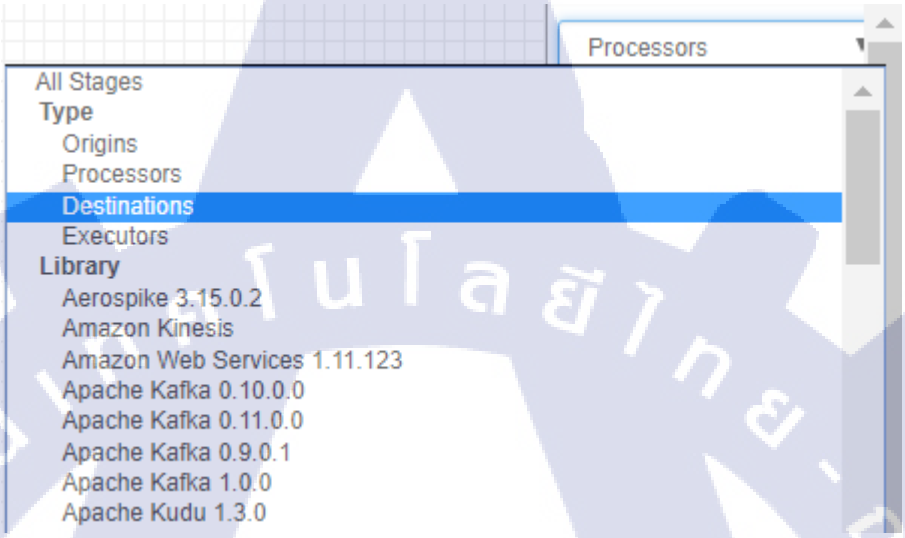
และเลือกหัวข้อData Format เลือก Avro



รูปที่ 4.76 ตั้งค่า Hive metadata

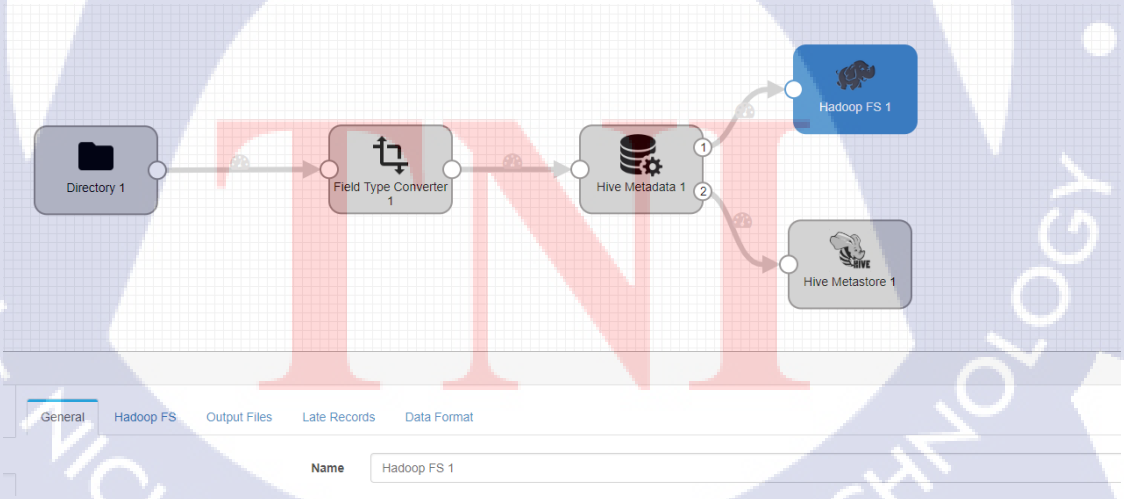
4) Destinations

Destinations เป็นส่วนปลายทางของข้อมูลว่าจะไปเก็บยังที่ใด ให้เลือก Stage Destinations เลือก stage Hadoop FS (write to a Hadoop fs) และเลือก Hive Metastore



รูปที่ 4.77 Stage destination

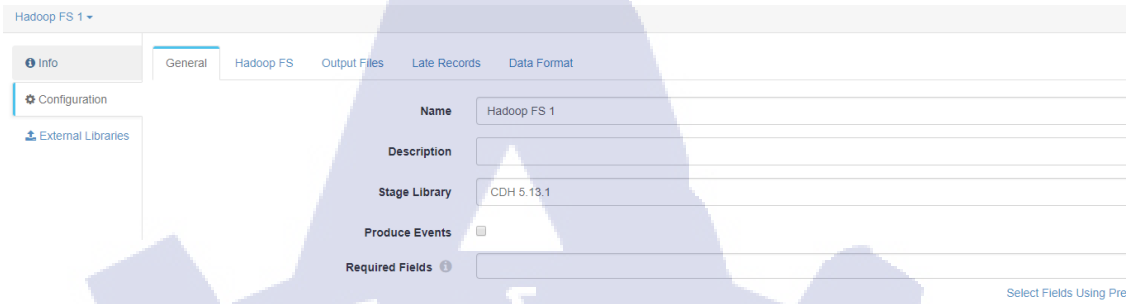
ให้ทำการเลือกเส้นทาง path ดังรูปที่ 4.78



รูปที่ 4.78 Destination HDFS และ Hive Metastore

เลือก Stage Hadoop fs เลือกหัวข้อ General

Stage Library : CDH x.xx.xx



รูปที่ 4.79 Hadoop fs general

หัวข้อ Hadoop FS

Hadoop FS URL : hdfs://localhost หรือใช้เป็น IP ก็ได้ในที่นี้ผู้จัดทำใส่ hdfs://cloudera.localhost หรือ hdfs://192.168.56.101

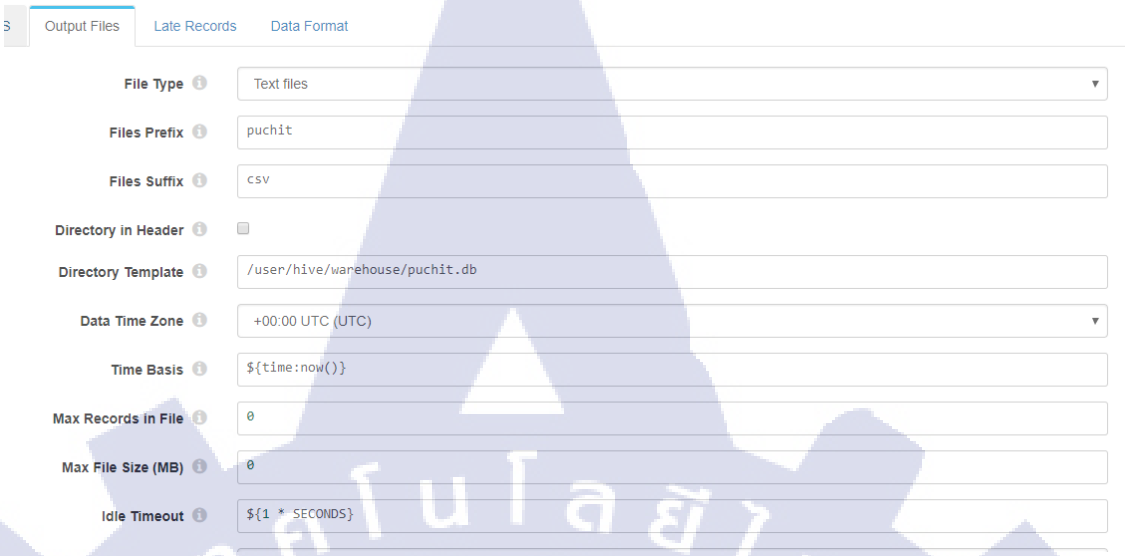
หัวข้อ Output Files

File Prefix : puchitid

File Suffix : csv

Directory template : /user/hive/warehouse/test.db (คือ hdfs path ที่เก็บ database ของผู้ใช้งาน)

Idel timeout : ${1 * SECONDS}



รูปที่ 4.80 Hadoop fs หัวข้อ Output files

หัวข้อ Data format

Data Format : Avro

Avro Schema Location : In Record Header

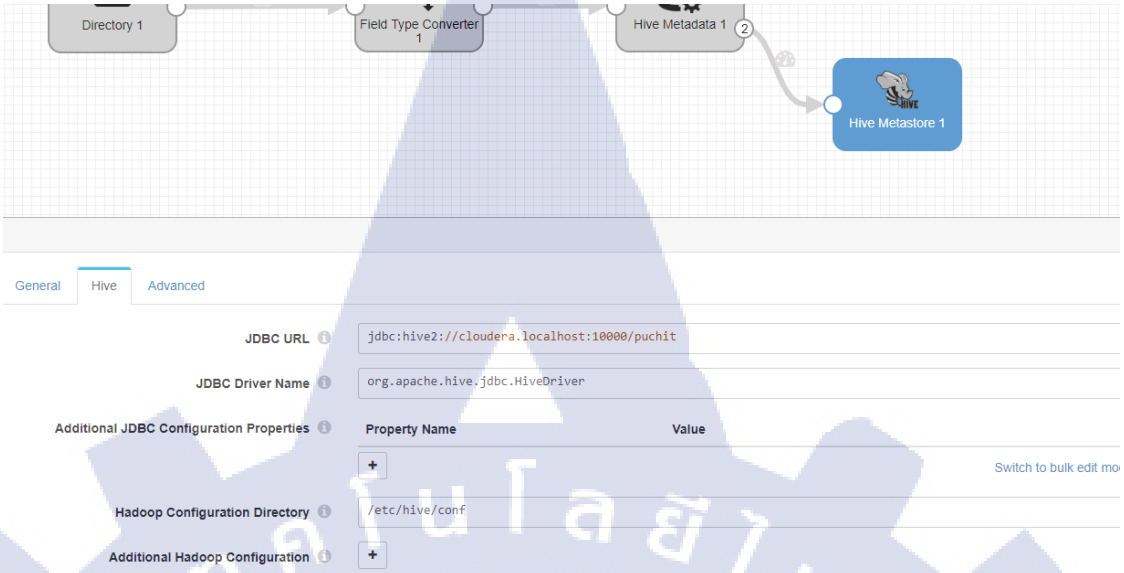
ต่อมาให้เลือก Stage Hive Metastore

หัวข้อ General

Stage Library : CDH x.xx.xx

หัวข้อ Hive

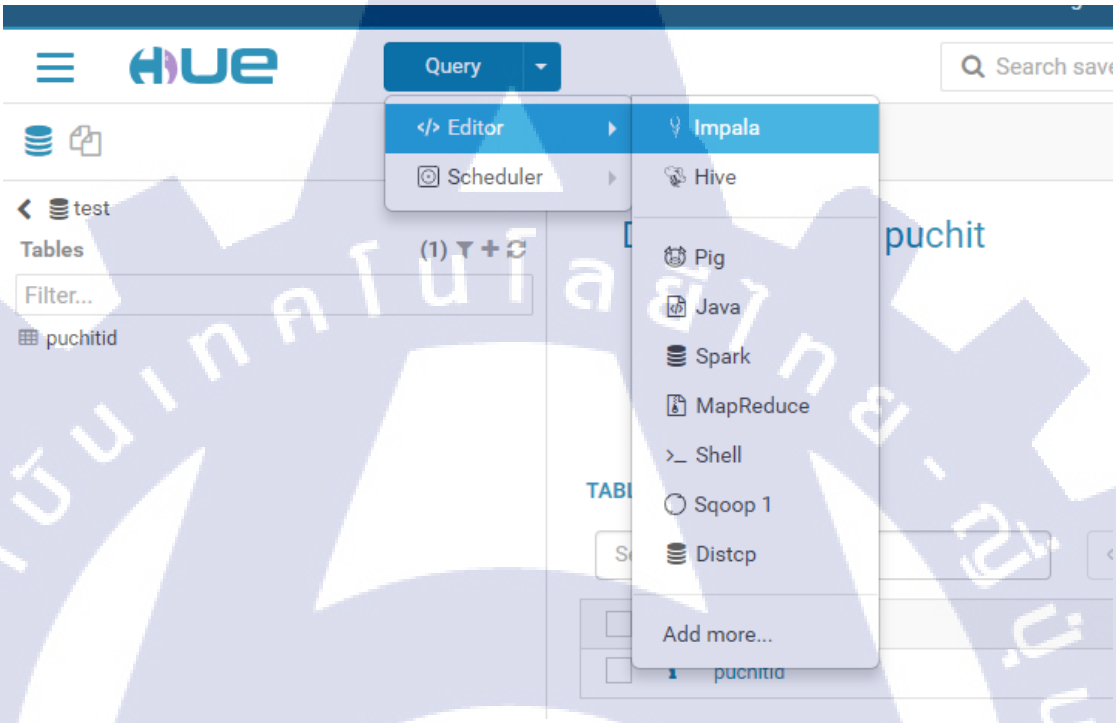
JDBC URL : jdbc:hive2://cloudera.localhost:10000/puchit หรือ จะใช้เป็น IPก็ได้ เช่น jdbc:hive2://192.168.56.101:10000/puchit



รูปที่ 4.81 Hive Metastore หัวข้อ hive

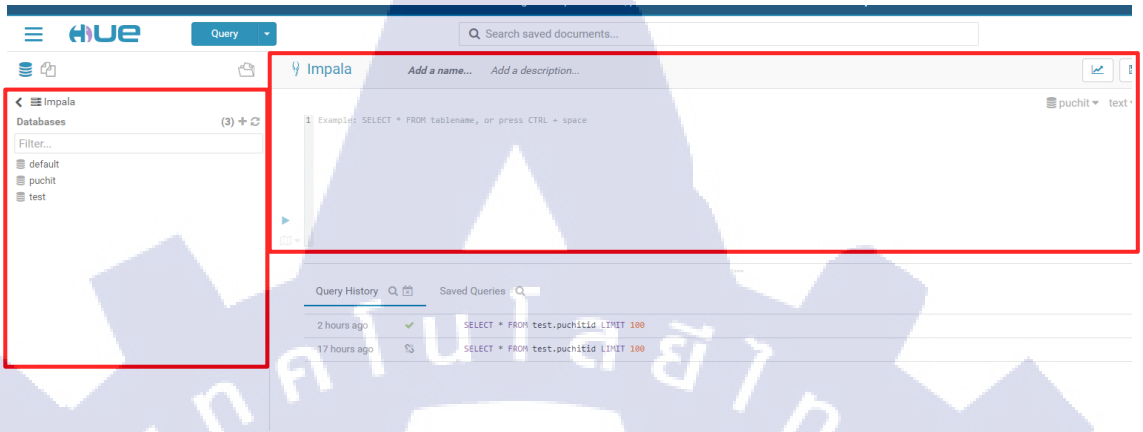
4.1.5 ทดสอบความถูกต้องหลังการโอนข้อมูลเข้า HDFS

เมื่อผู้จัดทำนำข้อมูลรูปแบบตารางเข้าก็ต้องทำการ Query ข้อมูลเพื่อทดสอบความถูกต้อง โดยการใช้ Impala ให้ไปที่ hue -> document -> editor -> Impala ดังรูปที่ 4.82



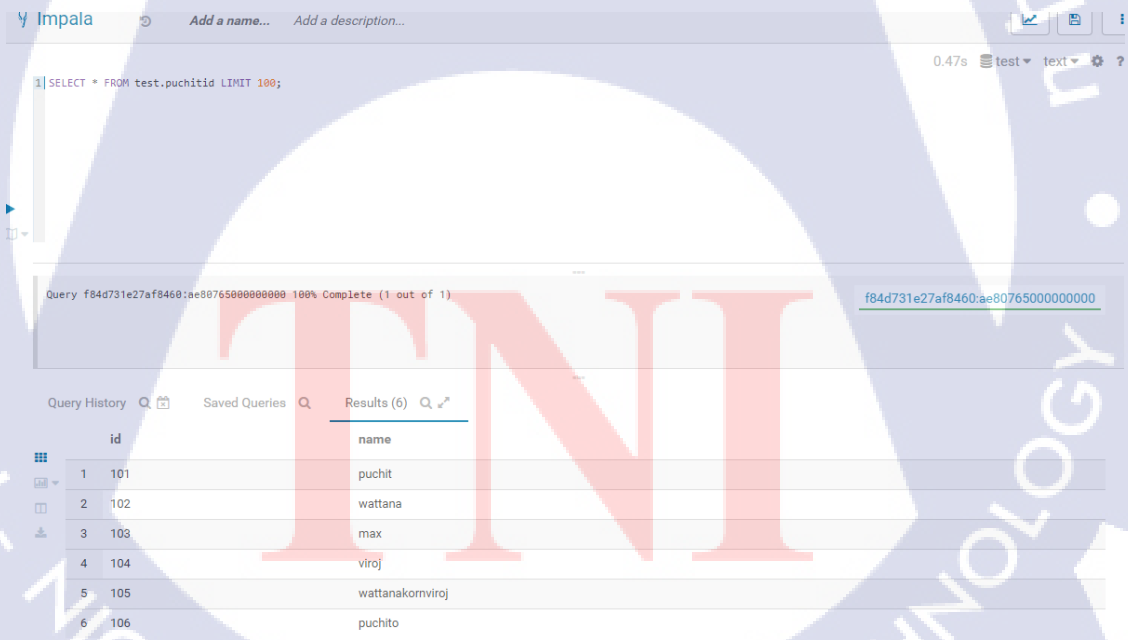
รูปที่ 4.82 การเข้า Impala

เมื่อเข้ามาแล้วจะได้หน้าดังรูปที่ 4.83 ช่องด้านซ้ายจะเป็น database สามารถกด คลิกเพื่อเข้าไปดู table ได้ส่วนช่องด้านขวาเป็นที่ใส่ Code SQL เพื่อ Query ข้อมูล



รูปที่ 4.83 Hue Impala

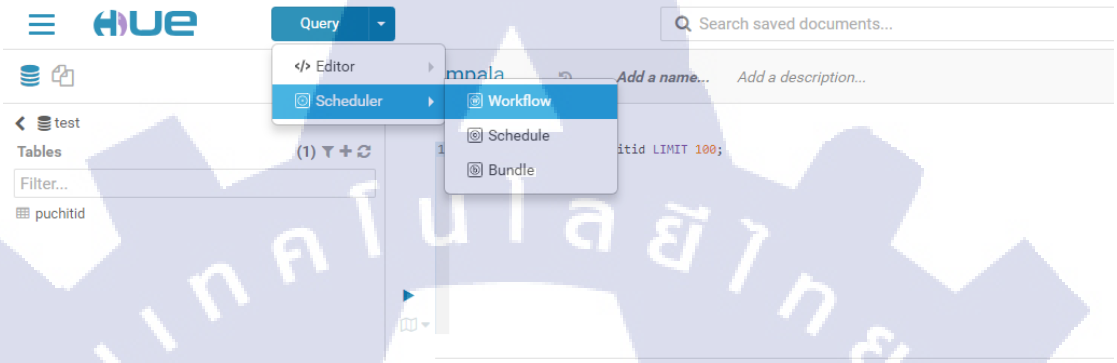
ตัวอย่างการใส่ code SQL ดังรูปที่ 4.84



รูปที่ 4.84 ตัวอย่างการใช้ code SQL

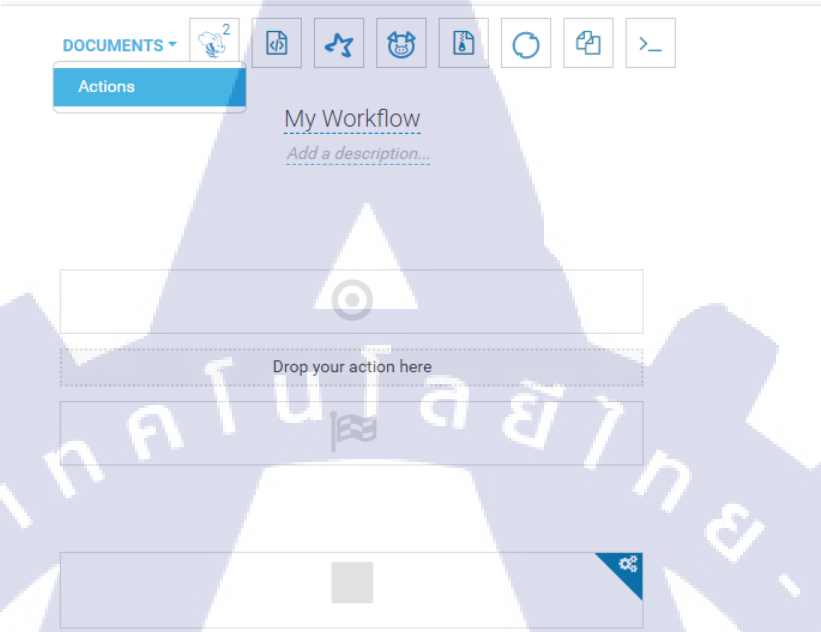
4.1.6 การตั้งเวลาในการทำงานให้ StreamSets

เมื่อเราต้องการให้ StreamSets ทำงานเวลาที่เราต้องการและปิดตามที่ต้องการเพื่อเป็นการลด performance เครื่อง โดยใช้เครื่อง Ozzie เพื่อrun Job schedule ให้ไปที่ Document -> Scheduler -> workflow ดังรูปที่ 4.85



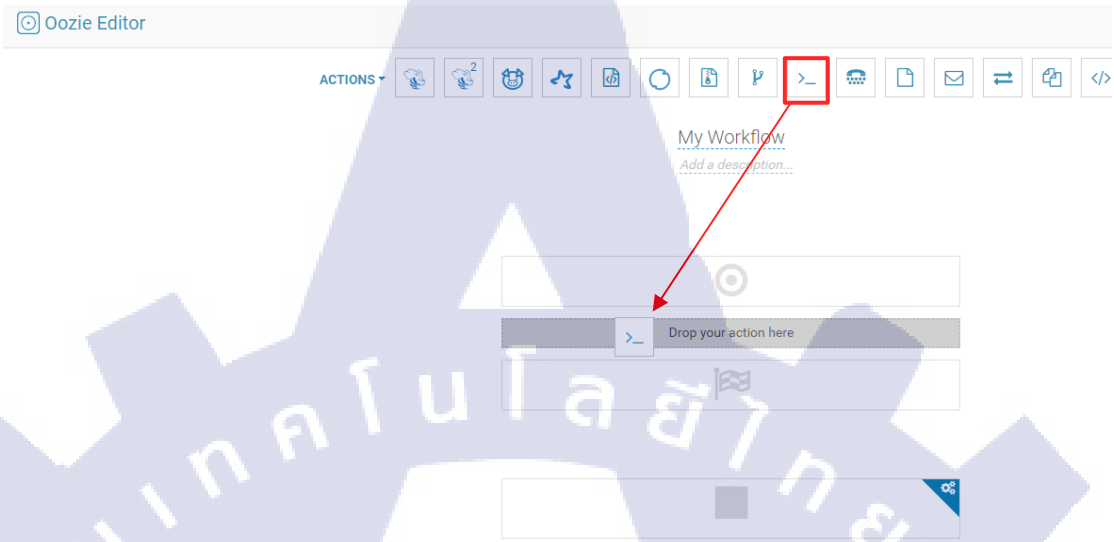
รูปที่ 4.85 เข้า Ozzie

เมื่อกดเข้ามาจะเห็นDocument กดเลือกแล้วเปลี่ยนเป็น Actionดังรูปที่ 4.86



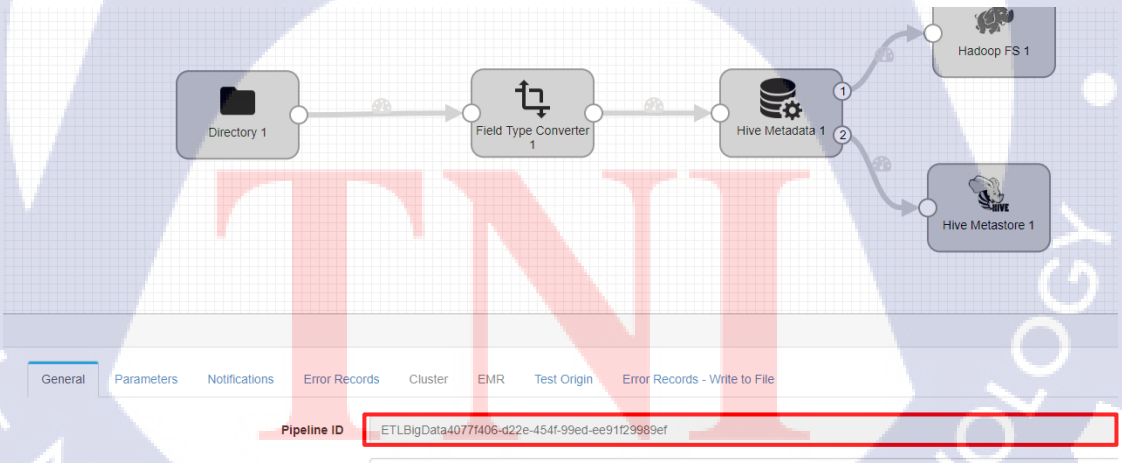
รูปที่ 4.86 Ozzie เปลี่ยน Document เป็น actions

เมื่อเลือกแล้วจะเห็นเป็นดังรูปที่ 4.87 ให้เลือก shell ลากไปที่ drop your action

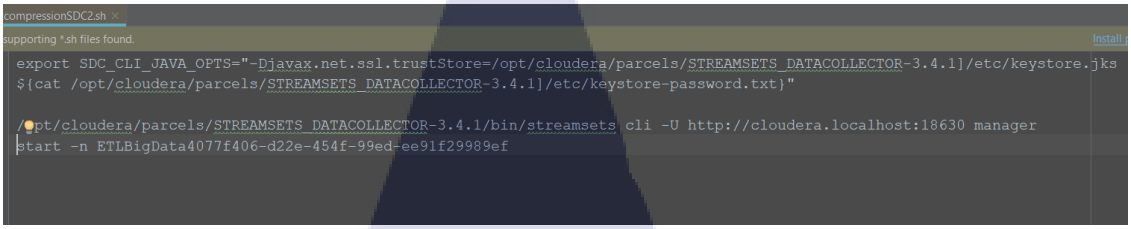


รูปที่ 4.87 การเลือก Action ที่ต้องการทำงาน

ให้เราเขียน Shell Script ขึ้นมาให้เราใช้คำสั่ง Start pipeline streamsets ขึ้นมา

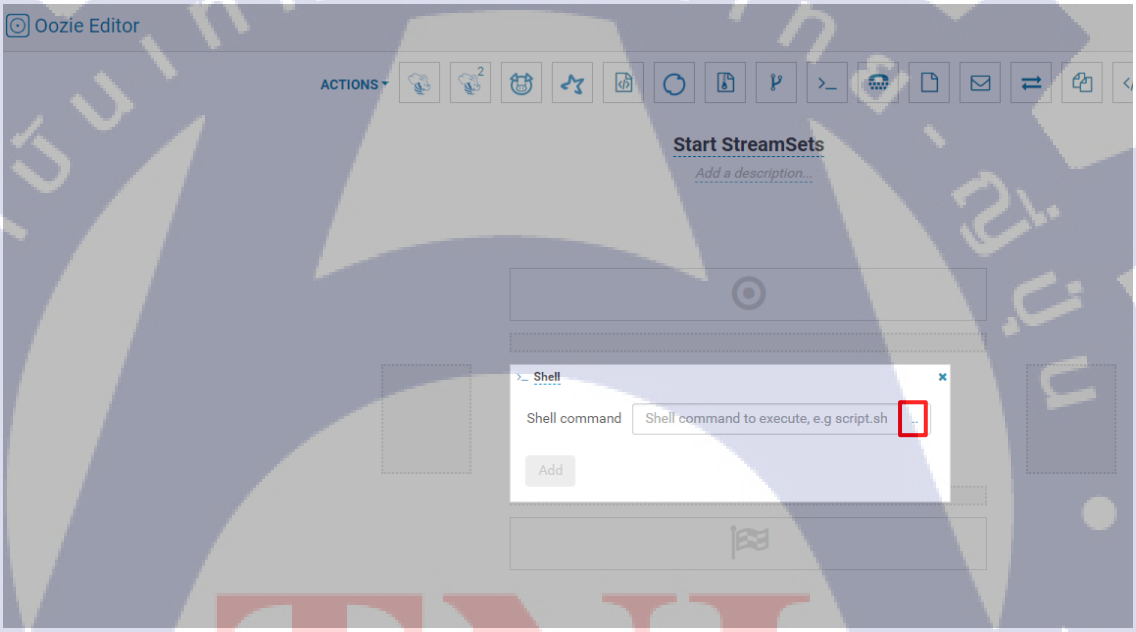


รูปที่ 4.88 Pipeline id



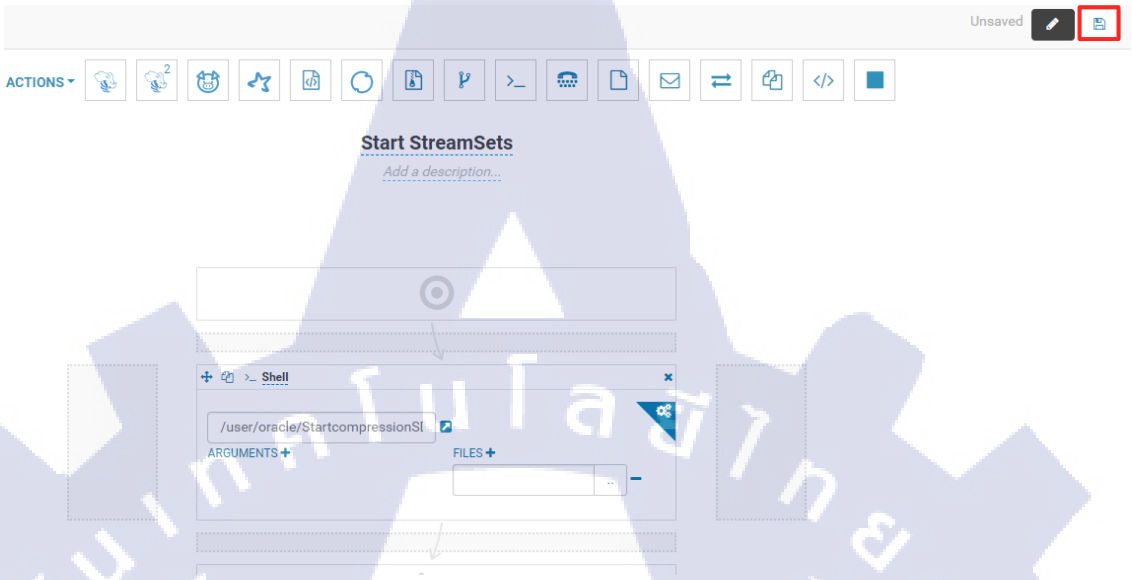
รูปที่ 4.89 Script การสั่งstart StreamSets

เมื่อได้ไฟล์ script แล้วให้ทำการ upload เข้า hdfs กลับไปหน้า hue ozzie เมื่อเรานำshell เข้ามาจะได้ดังรูปที่ 4.90 ให้เรากดเลือกปุ่ม .. ในกรอบสีแดง และกด add



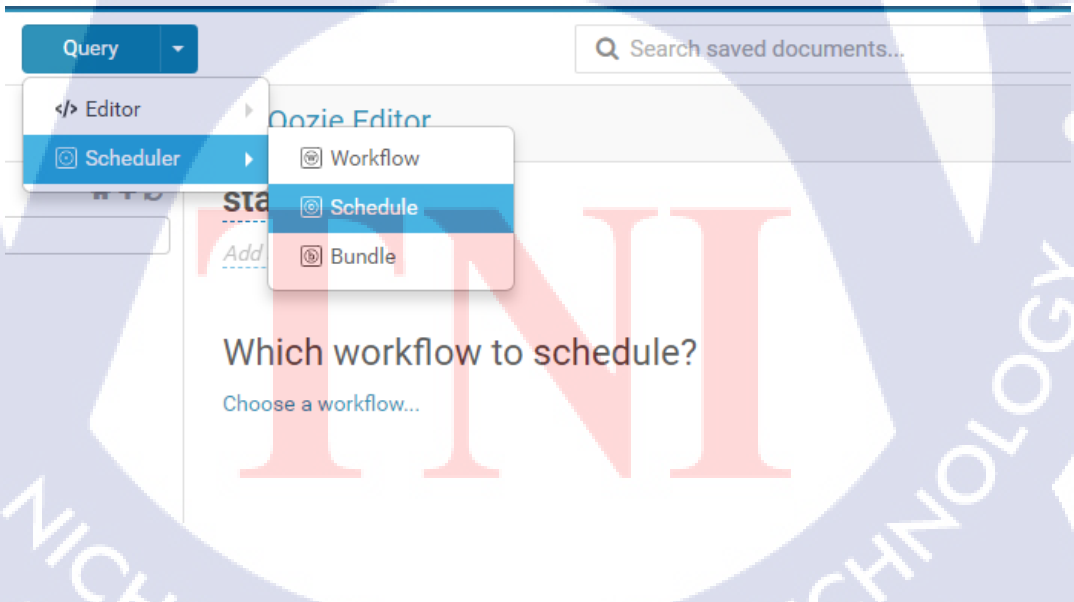
รูปที่ 4.90 Ozzie เมื่อนำshellมาว่าง

เมื่อได้ดังรูปที่ 4.91 แล้วให้กด รูปsave ในกรอบสีแดง



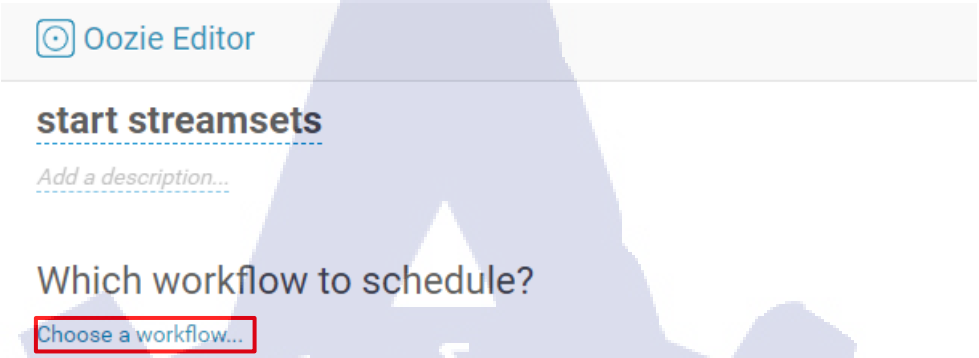
รูปที่ 4.91 บันทึก flow oozie

เมื่อsave แล้วให้เลือก document -> Scheduler->Schedule ดังรูปที่ 4.92



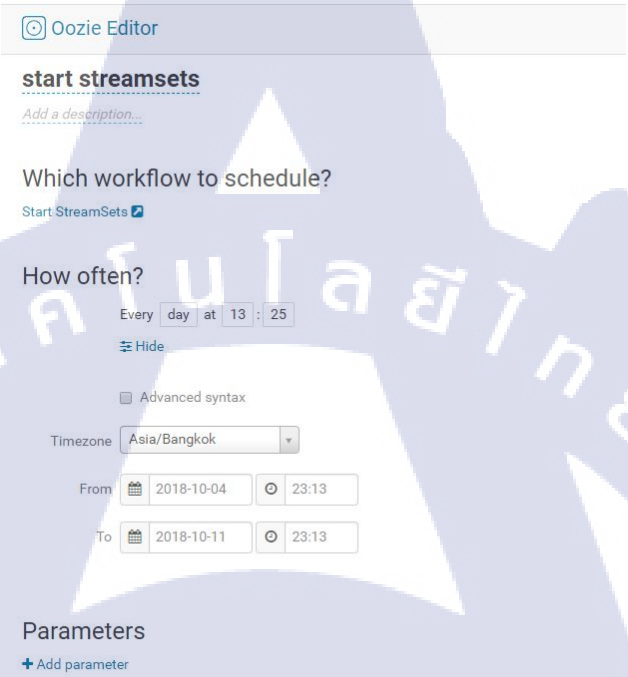
รูปที่ 4.92 Oozie Schedule

ให้ตั้งชื่อ schedule และเลือก Choose a workflow ดังรูปที่ 4.93



รูปที่ 4.93 เลือก schedule

เมื่อกดแล้วจะขึ้นpopupให้เราเลือก workflow เราก็เลือก Start StreamSets ที่สร้างเสร็จแล้ว เมื่อเลือกแล้วจะเป็นดังรูปที่ 4.94 เมื่อตั้งค่าเสร็จแล้วให้กด save และ submit เพื่อให้โปรแกรมทำงาน เท่านี้ก็จะสามารถตั้งให้StreamSets สามารถทำงานตามเวลาที่กำหนดได้



รูปที่ 4.94 ตั้งค่า schedule

บทที่ 5
บทสรุปและข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน

จากการที่ได้ศึกษาและทดลองทำ Big Data เพื่อการดึงและรวบรวมข้อมูลขนาดใหญ่เพื่อนำมาประมวลผลให้เกิดประโยชน์และมีประสิทธิภาพสูงสุดด้วยเครื่อง Bigdata ซึ่งได้ผลลัพธ์ดังนี้

5.1.1 สามารถติดตั้งระบบ Big Data และเครื่องมือต่างๆที่ใช้ในระบบได้สำเร็จ

5.1.2 สามารถนำข้อมูลเข้าฐานข้อมูลของBig Dataได้สำเร็จ

5.1.3 สามารถทำการดึงข้อมูลจากแหล่งข้อมูลมาปรับปรุงและนำเข้าฐานข้อมูล Big Data ได้สำเร็จ

5.1.4 สามารถกำหนดเวลาการทำงานโดยใช้เครื่องมือ Oozie ได้สำเร็จ

5.1.5 สามารถทำการทดสอบและได้ผลลัพธ์ตามที่คาดหวังไว้

จากผลลัพธ์ทั้งหมด สามารถสรุปได้ว่า การติดตั้งระบบ Cloudera Manager ซึ่งเป็นระบบจัดการเครื่องมือ Big Data สามารถตอบสนองต่อความต้องการของบริษัท และเครื่องมือ StreamSets สามารถทำการดึงข้อมูลจากแหล่งข้อมูล ปรับปรุงข้อมูล นำเข้าฐานข้อมูลของ Big Data ได้อย่างครบถ้วนตามที่ต้องการ พร้อมทั้งสามารถนำเทคโนโลยีทั้งหมดนี้ไปต่อยอดและพัฒนาในอนาคตได้

5.2 แนวทางการแก้ไขปัญหา

ปัญหาที่พบในระหว่างการศึกษาและติดตั้งระบบ Big Data นั้นส่วนแรกจะเป็นส่วนของการทำความเข้าใจในทฤษฎีและภาพรวมของโครงสร้าง Big Data เนื่องจากระบบของ Big Data มีโครงสร้างหลายขั้นตอนพอสมควร ซึ่งทำให้การทำความเข้าใจกระบวนการศึกษาเพื่อใช้งานเบื้องต้น มีความเข้าใจได้ยาก แนวทางการแก้ปัญหานี้คือ เรียนรู้ผ่านเว็บไซต์ที่มีผู้พัฒนาได้มาเขียนไว้ หรือศึกษาจากเอกสารหลักของเรื่องนั้นๆ ส่วนอีกปัญหาที่พบคือ การติดตั้งโปรแกรม Cloudera Manager นั้นหากตั้งค่าของโปรแกรมต่างๆไม่ถูกต้อง หรือไม่เหมาะสมแก่การติดตั้งในส่วนต่อไปค่อนข้างมีความยุ่งยาก และอาจเกิดปัญหาการเข้าใช้งานไม่ได้ ซึ่งแนวทางการแก้ปัญหาคือ ศึกษาว่า

Cloudera Manager จะต้องการสิ่งใดบ้างเพื่อที่จะได้วางแผนการติดตั้งระบบได้ และในการตั้งค่าต่างๆควรรอบคอบละเอียดที่สุด

5.3 ข้อเสนอแนะจากการดำเนินงาน

5.3.1 ควรมีพื้นฐานในการใช้ระบบปฏิบัติการ Linux ทางด้านการใช้คำสั่งและการตั้งค่าระบบพื้นฐานของเครื่อง

5.3.2 ควรมีพื้นฐานและความเข้าใจในการติดตั้ง Package เนื่องจาก Cloudera Manager ต้องการ Package ก่อนการติดตั้งหลาย Package ด้วยกัน

5.3.3 ควรมีพื้นฐานการใช้งานและติดตั้ง MySQL เนื่องจากเครื่อง Big Data บ้างเครื่องมือจำเป็นต้องมีที่เก็บข้อมูล

5.3.4 ในการทำงานเป็นต้องหาข้อมูลเพื่อเตรียมความพร้อมก่อนลงมือและต้องวางแผนก่อนการทำงานเสมอ

เอกสารอ้างอิง

Installation Path B - Installation Using Cloudera Manager Parcels or Packages [Online],Avilable : https://www.cloudera.com/documentation/enterprise/5-14-x/topics/cm_ig_install_path_b.html [1 ตุลาคม 2561]

Cloudera Manager and Managed Service Datastores [Online],Available : https://www.cloudera.com/documentation/enterprise/5-14-x/topics /cm_ig_installing_configuring_dbs.html #cmig_topic_5 [3 ตุลาคม 2561]

Hadoop: Setting up a Single Node Cluster [Online],Available : http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html [5 ตุลาคม 2561]

Hue User Guide [Online],Available : http://cloudera.github.io/hue/latest/user-guide/user-guide.html [8 ตุลาคม 2561]

StreamSets Documentation [Online],Available : https://streamsets.com/documentation-page [8 ตุลาคม 2561]

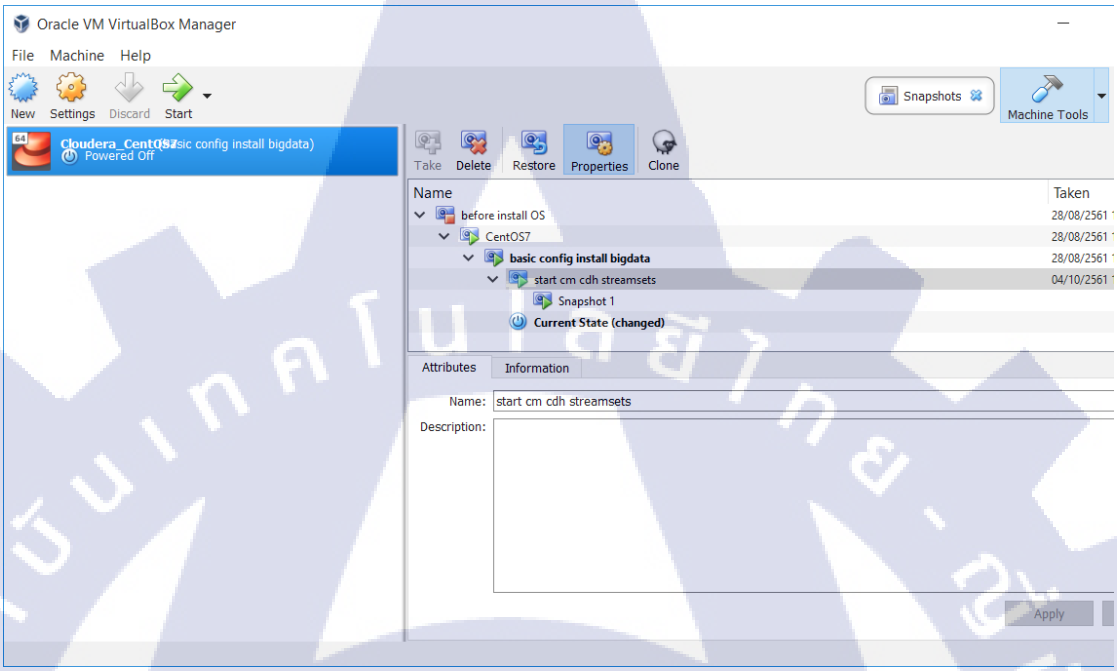
ภาคผนวก

ภาคผนวก ก.

การเตรียม Virtual machine โดยใช้ Oracle VM VirtualBox

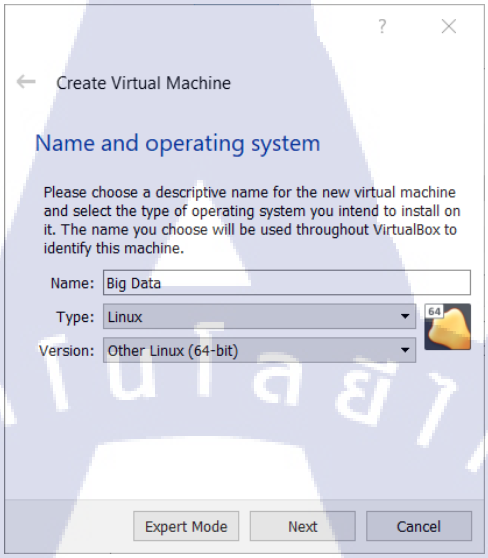
การเตรียม Virtual Machine โดยใช้ Oracle VMware VirtualBox

- เข้าสู่หน้า VMware



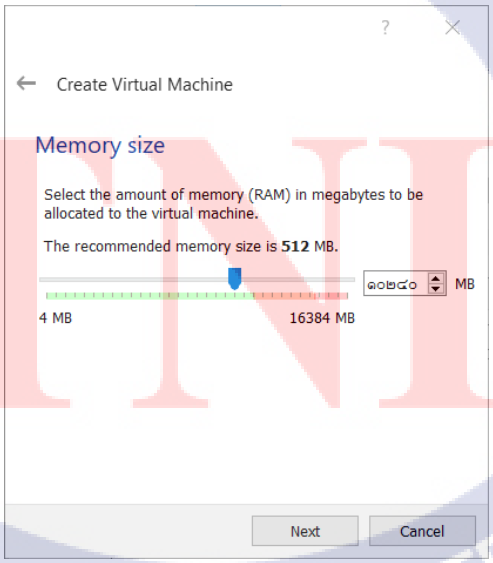
รูปที่ ก.1 การติดตั้ง VMware (1)

- เลือก New และตั้งชื่อ type : linux version : Other Linux (64-bit)



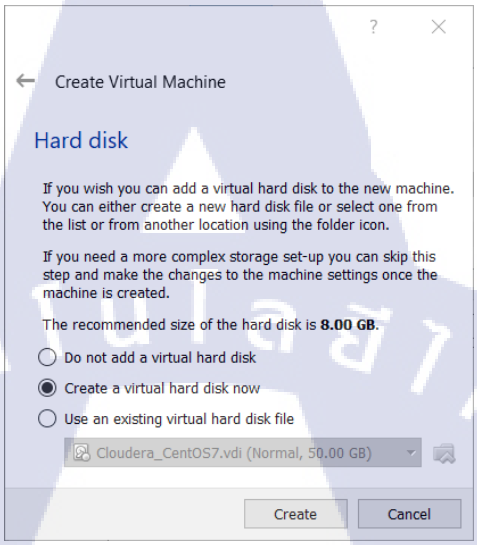
รูปที่ ก.2 การติดตั้ง VMware (2)

- เลือดขนาด RAM การติดตั้ง Cloudera อย่างน้อยต้องมี 32 GB แต่เนื่องจากเอาให้ใช้ได้เบื้องต้นจึงตั้ง 10240 mb



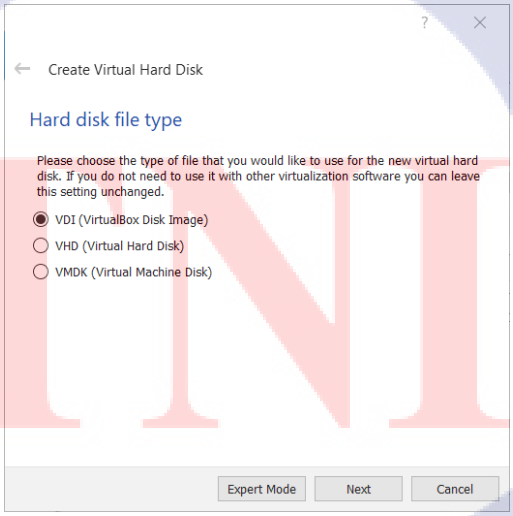
รูปที่ ก.3 การติดตั้ง VMware (3)

- เลือก Create a virtual gard disk now



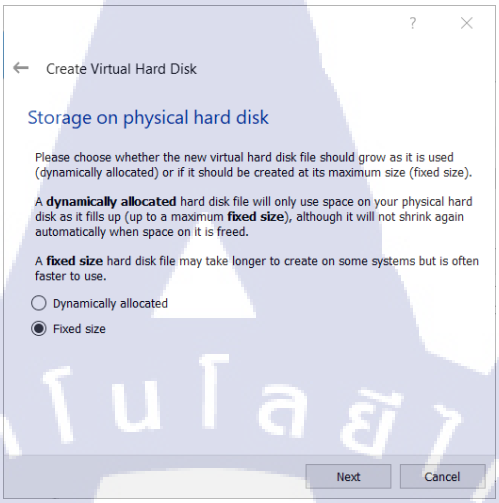
รูปที่ ก.4 การติดตั้ง VMware (4)

- เลือก VDI



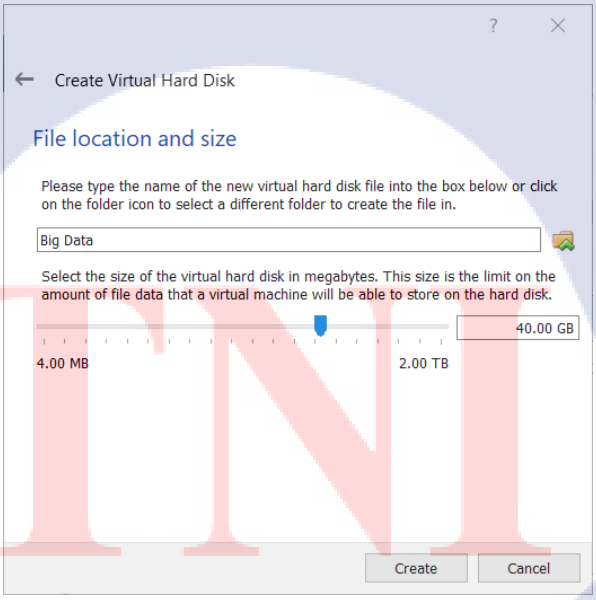
รูปที่ ก.5 การติดตั้ง VMware (5)

- เลือก Fixed size



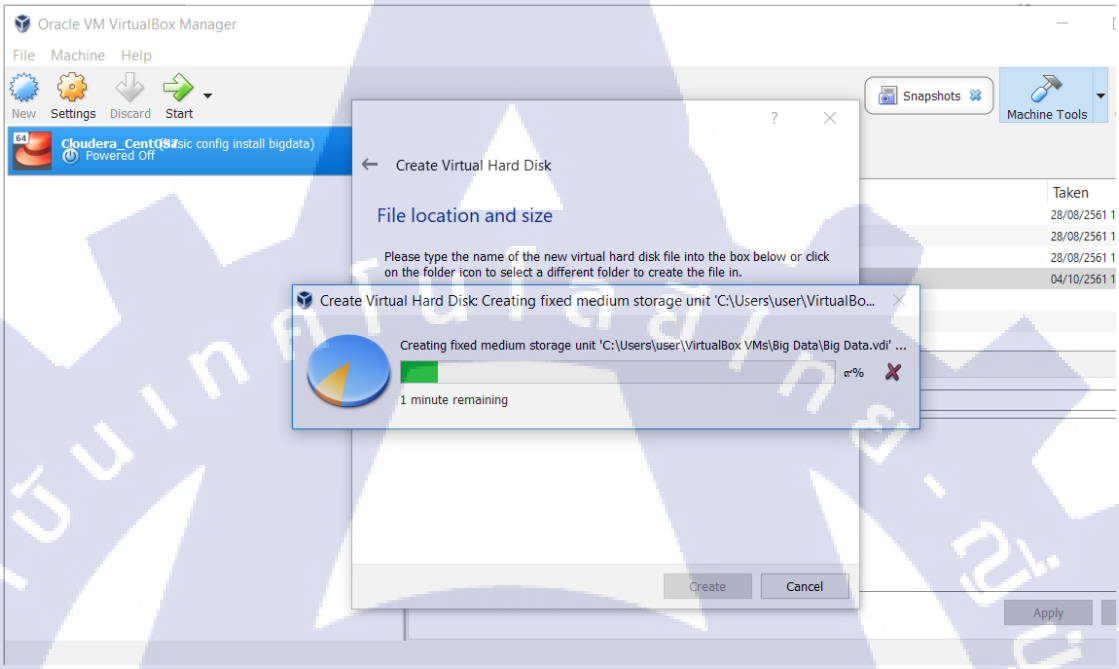
รูปที่ ก.6 การติดตั้ง VMware (6)

- เลือก size hard disk 40 GB



รูปที่ ก.7 การติดตั้ง VMware (7)

- รอการVMware สร้างที่เก็บข้อมูล



รูปที่ ก.8 การติดตั้ง VMware (8)

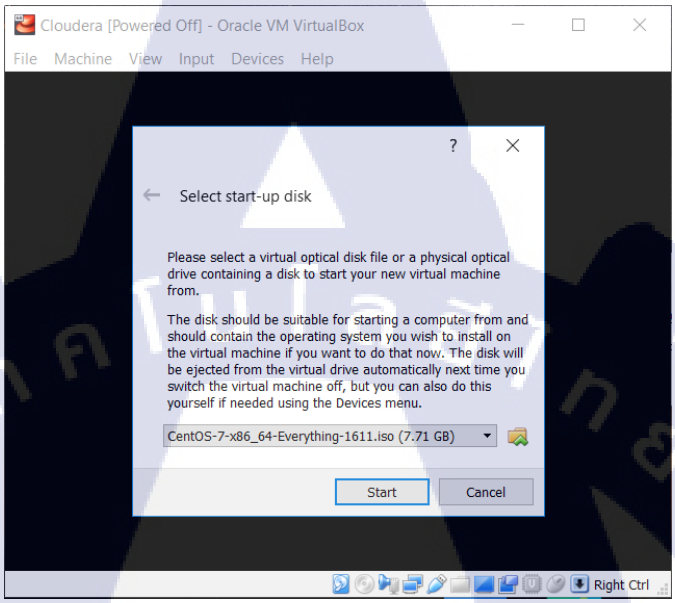
- เมื่อเราได้ VMware แล้วกดกด Start เพื่อเข้าโปรแกรม

ภาคผนวก ข.

การติดตั้ง CentOS 7 Linux

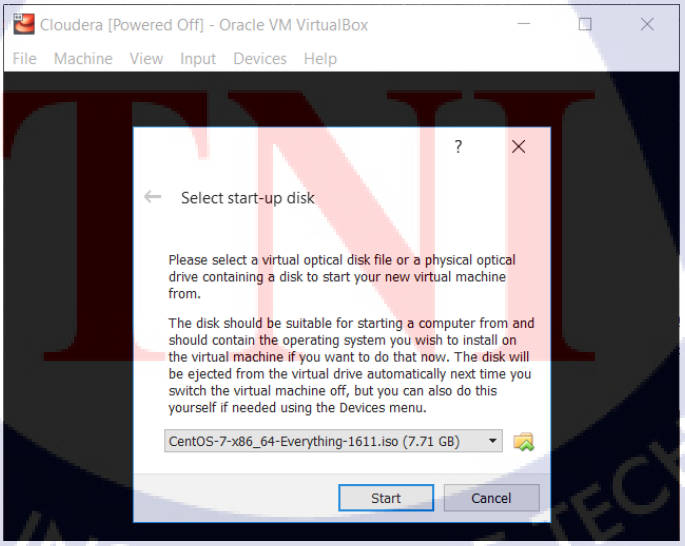
การติดตั้ง CentOS 7 Linux

- เลือก CentOS 7 iso ไฟล์ที่เตรียมไว้และกด Start



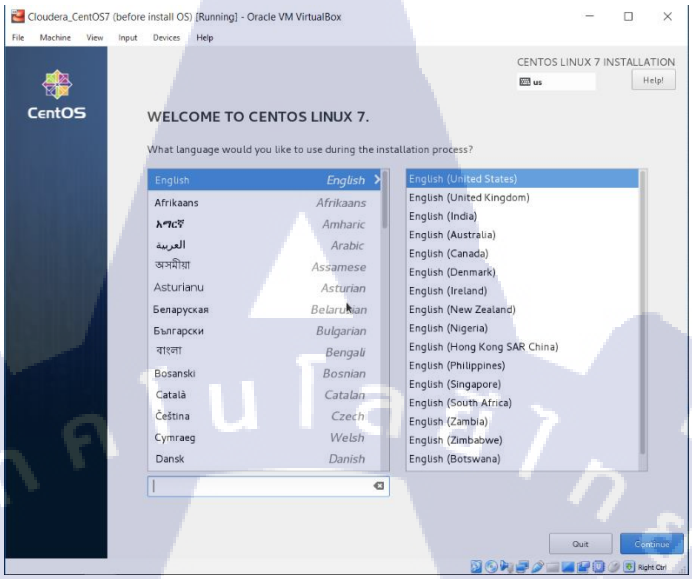
รูปที่ ข.1 การติดตั้ง Linux (1)

- เลือก install CentOS 7



รูปที่ ข.2 การติดตั้ง Linux (2)

- เลือกภาษาอังกฤษ (USA)

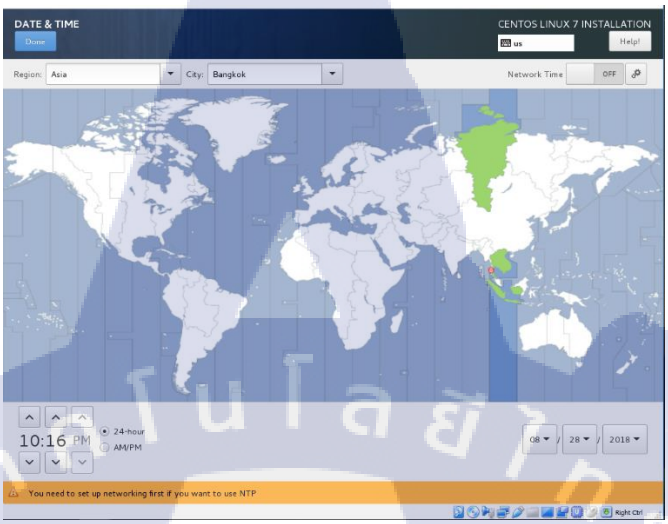


รูปที่ ข.3 การติดตั้ง Linux (3)

- เมื่อขึ้นหน้านี้ให้เลือก Date & Time

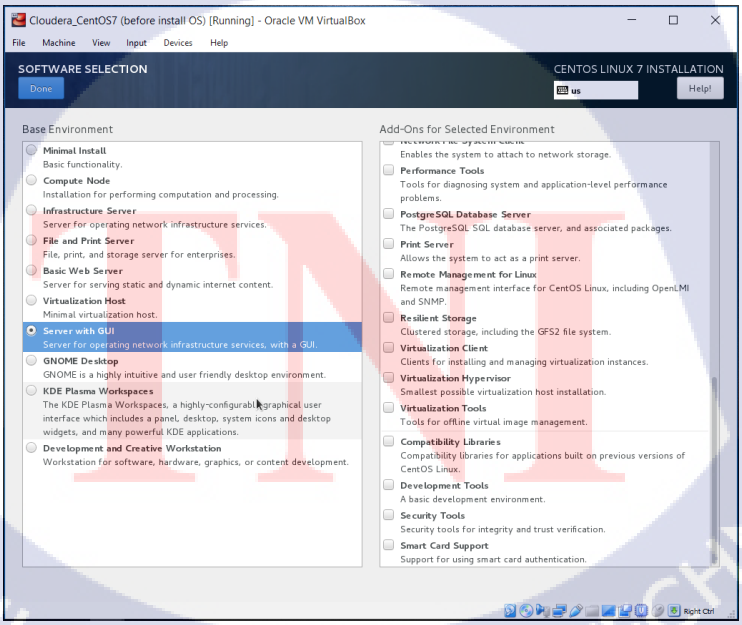
รูปที่ ข.4 การติดตั้ง Linux (4)

- เลือก Asia / Bangkok และกด Done



รูปที่ ข.5 การติดตั้ง Linux (5)

- กลับมาหน้าหลักแล้วเลือก Software Selection ให้เลือก Server with GUI และกด Done

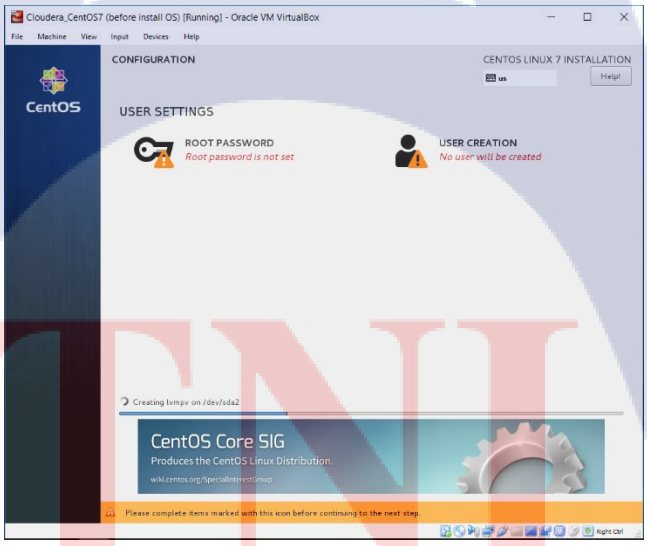


รูปที่ ข.6 การติดตั้ง Linux (6)

- กลับมาหน้าหลักเลือก Installation Destination เพื่อจัดการเรื่องdisk

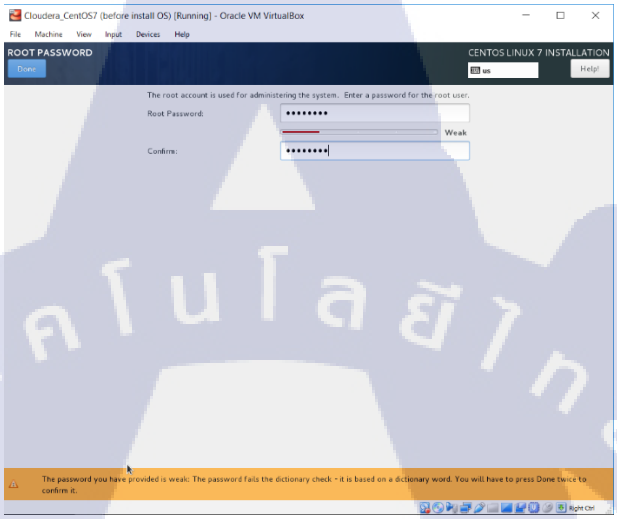
รูปที่ ข.7 การติดตั้ง Linux (7)

- กลับมาหน้าหลักจะสามาถกด Begin Installation ได้และจะเริ่ม install



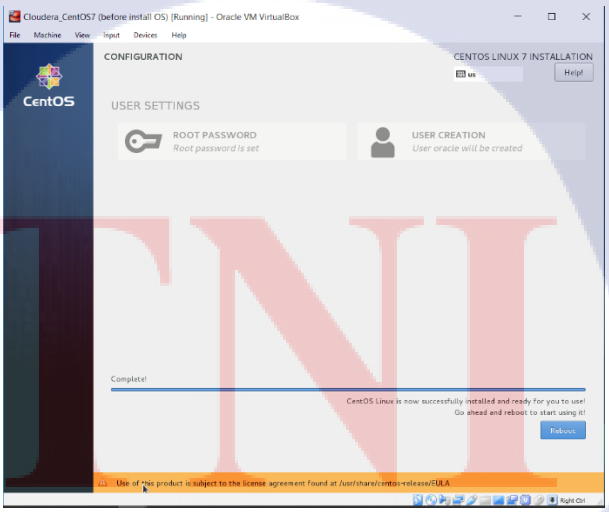
รูปที่ ข.8 การติดตั้ง Linux (8)

- ให้เลือก root และตั้งpassword



รูปที่ ข.9 การติดตั้ง Linux (9)

- กด Reboot เครื่องก็จะเป็นการติดตั้ง CentOS เสร็จเรียบร้อย



รูปที่ ข.10 การติดตั้ง Linux (10)

ประวัติผู้จัดทำโครงงาน

ชื่อ – สกุล นายภูชิสส์ วัฒนกรวิโรจน์

วัน เดือน ปีเกิด 26 ธันวาคม 2539

ประวัติการศึกษา

ระดับประถมศึกษา โรงอุดมศึกษาลาดพร้าว

ระดับมัธยมศึกษาตอนต้น โรงเรียนอุดมศึกษาลาดพร้าว

ระดับมัธยมศึกษาตอนปลาย โรงเรียนมัธยมวัดบึงทองหลาง

ระดับอุดมศึกษา คณะเทคโนโลยีสารสนเทศ สาขาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีไทย – ญี่ปุ่น

ทุนการศึกษา - ไม่มี -

ประวัติการฝึกอบรม Training Pre-Cooperative Education โครงการสหกิจฯบริษัทเอ-โฮสต์ จำกัด

ผลงานที่ได้รับการตีพิมพ์ - ไม่มี -