

การดึงและรวบรวมข้อมูลขนาดใหญ่เพื่อนำมาประมวลผลให้เกิดประโยชน์ และมีประสิทธิภาพสูงสุดด้วยเครื่องมือ Big Data **BIG DATA GATHERING FOR THE MOST EFFECTIVE ANALYSIS BY USING BIG DATA TOOLS**

นาย ภูชิสส์ วัฒนกรวิโรจน์

10

โครงงานสหกิจ<mark>ศึก</mark>ษานี้เป็นส่<mark>ว</mark>นหนึ่<mark>งของกา</mark>รศึกษ<mark>าตา</mark>มหลักสูตร ้ปริญญาวิทย<mark>าศาส</mark>ตรบัณฑิ<mark>ต</mark> สาขาวิ<mark>ชาเท</mark>คโนโล<mark>ยีสา</mark>รสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยี ไทย-ญี่ปุ่น W.M. 2001 WSTITUTE OF

การดึงและรวบรวมข้อมูลขนาดใหญ่เพื่อนำมาประมวลผลให้เกิดประโยชน์ และมีประสิทธิภาพสูงสุดด้วยเครื่องมือ 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

(อาจารย<mark>์ โอพ</mark>าร รื่นชื่น)

......อาจ<mark>ารย์ที่</mark>ปรึกษา

(อาจารย<mark>์ สลิล</mark>า ชีวกิดากา<mark>ร</mark>)

.....ประธานสหกิจศึกษาสาขาวิชา (อาจารย์ สลิลา ชีวกิดาการ)

_________________ลิขสิทธิ์ของสถาบันเทคโนโลยีไทย – ญี่ปุ่น

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

บทสรุป

ในการสหกิจศึกษาได้รับมอบหมายในตำแหน่ง Assistant Technical Consultant ให้ช่วยใน การสร้างระบบ Big Data ให้ บริษัท เอ - โฮสต์ จำกัด ที่มีวัตถุประสงค์ประสงค์เพื่อ ต้องการระบบที่ สามารถจัดการDestributed file system ได้ และเครื่องมือต่างๆของ Big Data ที่สามารถใช้งานได้ อย่างมีประสิทธิภาพสูงสุด

10

การติดตั้งระบบ Big Data นั้นได้ใช้เครื่องมือ Cloudera Manager ที่เป็นตัวจัดการให้ระบบ Big Data นั้นอยู่ในสภาพแวดล้อมเดียวกันเพื่อให้ Hadoop Destributed File System ซึ่งเป็น ฐานข้อมูลของ Big Data <mark>นั้น</mark> สามารถนำไปใช้โดยเครื่องมือ Big Data ได้อย่างมีประโยชน์และ ประสิทธิภาพสูงสุด

Diagram Big Data



กิตติกรรมประกาศ

ในการที่ข้าพเจ้าได้มาสหกิจศึกษา ณ บริษัท เอ-โฮสต์ จำกัด ตั้งแต่วันที่ 4 มิถุนายน พ.ศ. 2561 ถึงวันที่ 28 กันยายน พ.ศ. 2561 ได้ทำให้ข้าพเจ้าได้เรียนรู้ประสบการณ์ต่าง ๆ ความรู้จากการ ทำงานจริง ซึ่งมีค่าอย่างเป็นอย่างมาก และส่งผลให้ข้าพเจ้าสามารถนำสิ่งต่าง ๆ ที่ได้จากการมาสห กิจ เหล่านั้นมาใช้พัฒนาทักษะของตนเอง สำหรับรายงานการปฏิบัติงานสหกิจศึกษาในครั้งนี้ สามารถสำเร็จลุล่วงได้ด้วยดีจากความร่วมมือและการสนับสนุนจากหลายฝ่ายดังนี้

- 1. คุณบุญประสิทธิ์ ตั้งชัยสุข
- 2. คุณสุชัย เย็นฤดี
- 3. คุณเปรมสินี พิพัฒน์โกศล
- 4. คุณพิชานน จะเรียมพันธ์
- 5. คุณนฤตยา สมศริกุล
 6. คุณอภิสิทธิ์ แซ่ตั้ง

(6)

(Chief Executive Officer) (Senior Vice President) (HR Manager) (Manager) (Programmer) (Programmer)

และบุคลากรท่านอื่น ๆ ที่ไม่ได้กล่าวนามทุกท่านที่ได้ให้คำแนะนำช่วยเหลือในการจัดทำ รายงานข้าพเจ้าขอขอบพระคุณ ผู้มีส่วนเกี่ยวข้องทุกท่านที่มีส่วนร่วมในการให้ข้อมูลเป็นที่ปรึกษา ในการทำรายงานฉบับนี้จนเสร็จสมบูรณ์ ตลอดจนให้การดูแลและให้ความเข้าใจเกี่ยวกับชีวิตของ การทำงานจริงข้าพเจ้าขอขอบพระคุณไว้ ณ ที่นี้

> นาย ภูชิสส์ วัฒนกรวิโรจน์ ผู้จัดทำ

สารบัญ

บทสรุป		
กิตติกรรมประกาศ		
สารบัญ		
สารบัญรูป		
สารบัญตาราง		
บทที่ 🦳 🗛 💧	ula	87

บทที่

1

2

บข	າนຳ		1
1	.1	ชื่อและที่ตั้งของสถานประกอบการ	.1
1	.2	ลักษณะธุรกิจของสถานประกอบการ หรือการให้บริการหลักขององค์กร	.2
1	.3	รูปแบบการจัดองค์กรและการบริหารองค์กร	.4
1	.4	ตำแหน่งและหน้าที่งานที่นักศึกษาได้รับมอบหมาย	.5
1	.5	พนักงานที่ปรึกษา และ ตำแหน่งของพนักงานที่ปรึกษา	.5
1	.6	ระยะเวลาที่ปฏิบัติงาน	.5
1	.7	ที่มาและความสำคัญของปัญหา	.5
1	.8	วัตถุประสงค์หรือจุดมุ่งหมายของโครงงาน	.6
1	.9	ผลที่คาคว่าจะได้รับจากการปฏิบัติงานหรือโครงงานที่ได้รับมอบหมาย	.6
1	.10	นิยามศัพท์เฉพาะ	.6
ทธ	ษฎีเ	และเทคโนโลยีที่ <mark>ใช้ใน</mark> การปฏิบัติง <mark>า</mark> น ()	7
2	1 78	ะบบปัญญาธุรกิจ <mark> (Bu</mark> siness Intell <mark>i</mark> gence: BI)	.7
	2.	1.1 ลักษณะสำค <mark>ัญขอ</mark> งระบบ BI <mark>ค</mark> ือ	.8
	2.	1.2 กระบวนการทำงานของ BI	.8
	2.	1.3 ประโยชน์ของ BI1	15
2	.2 ข้า	อมูลขนาดใหญ่ (Bigdata)1	16
2	.3 เท	าคโนโลยีที่ใช้ในการปฏิบัติงาน	17

ก

ค

٩

R

IJ

2.3.2 โปรแกรม Cloudera	
2.3.3 Apache Hadoop	19
2.3.4 โปรแกรม Hue (Hadoop User Experience)	20
2.3.5 Apache Hive	21
2.3.6 โปรแกรม Impala	21
2.3.7 โปรแกรม Ozzie	22
2.3.8 โปรแกรม Kafka	23
3 แผนงานการปฏิบัติงานและขั้นตอนการดำเนินงาน	25
3.1 แผนงานการฝึกงาน	
3.2 รายละเอียดที่นักศึกษาปฏิบัติในการฝึกงาน	26
3.3 ขั้นตอนการคำเนินงานที่นักศึกษาปฏิบัติงาน	
3.3.1 ศึกษา Business Intelligence และ Big Data	26
3.3.2 ศึกษากระบวนการ ETL	
3.3.3 ศึกษาการใช้และทำความเข้าใจ StreamSets	
3.3.4 ศึกษาเครื่องมือและทำความเข้าใจ Hue	
3.3.5 ศึกษาเครื่องมือและทำความเข้าใจ Hive.Impala	
3.3.6 ศึกษาเครื่องมือ Ozzie	
3.3.7 ศึกษาระบบ Kafka	2.7
4 สรปผลการดำเนิ <mark>นงาน การวิเคราะ</mark> ห์และสรปผลต่าง ๆ	28
4.1 ขั้นตอนและผล <mark>การค</mark> ำเนินงาน	
4.1.1 การติดตั้งโป <mark>รแก</mark> รม Cloudera Manager 114 Centos 7	28
4.1.2 ติดตั้งโปรแกรม StreamSets ถึงบน Cloudera Manager	59
4 1 3 การบำข้อมูลเข้า Hadoon (Hadoon Distrubute File System) โดยการใช้ Hue	68
4 1 4 การทำ FTL ด้วย StreamSets	
4 1 5 ทดสดบคาาบุกกต้องหลังการโอบข้อบลเข้า HDFS	۳ /
4 1 6 การตั้งเาลาใบการทำงานให้ StreemSate	
	91

5 บทสรุปและข้อเสนอแนะ

98

จ

5.1	สรุปผลการคำเนินงาน		
5.2	แนวทางการแก้ไขปัญหา		
53	ข้อเสบอบบะจากการดำเบิบ	งาาเ	99
5.5		N 1 Po	

เอกสารอ้างอิง

ภาคผนวก			101
ก. การเตรียม Virtual ma	chine โดยใช้ Oracle VM	VirtualBox	
ข. การติดตั้ง CentOS 7 I	_inux		

ประวัติผู้จัดทำโครงงาน

114

100

ฉ

WSTITUTE OF TECH

สารบัญรูป

¥

รูปที่ 4.3 การเปลี่ยน selinux เป็น Disabled	30
รูปที่ 4.4 การตั้งค่า swappiness	
รูปที่ 4.5 การตั้งค่า rc.local	
รูปที่ 4.6 การใช้คำสั่งtar ไฟล์ไปยัง directoryที่ต้องการ	
รูปที่ 4.7 การใส่ path ให้กับที่อยู่ของJDK	
รูปที่ 4.8 ตัวอย่างการแก้ path	33
รูปที่ 4.9 การเข้า mariadb ด้วย command line	
รูปที่ 4.10 ตัวอย่างการเปลี่ยนรหัสผ่าน	
รูปที่ 4.11 ตัวอย่าง syntax	35
รูปที่ 4.12 หน้าWeb UI cloudera Manager	
รูปที่ 4.13 User License	
รูปที่ 4.14 หน้าเลือกรูปแบบการทำงาน	38
รูปที่ 4.15 thank you for choosing Cloudera Manager and CDH	39
รูปที่ 4.16 ตั้งค่า hosts	
รูปที่ 4.17 ตั้งค่า Host เมื่อกคsearch	41
sun 4 19 ensuane Demositerry	N
มู่มที่ 4.18 การแขก Repository	
รูปที่ 4.19 หน้า Accept JDk License	
รูปที่ 4.19 หน้า Accept JDk Licenseรูปที่ 4.20 หน้า Single user mode	
รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials	
รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials รูปที่ 4.22 หน้า Install agent	
รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials รูปที่ 4.22 หน้า Install agent รูปที่ 4.23 Install agent success	42 43 44 45 46 47
รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials รูปที่ 4.22 หน้า Install agent รูปที่ 4.23 Install agent success รูปที่ 4.24 หน้า Detecting CDH Versions	42 43 44 45 46 47 48
รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials รูปที่ 4.22 หน้า Install agent รูปที่ 4.23 Install agent success รูปที่ 4.24 หน้า Detecting CDH Versions รูปที่ 4.25 หน้า Inspect hosts for correctness	42 43 44 45 46 46 47 48 49
รูปที่ 4.18 ทามเอก Repository รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials รูปที่ 4.22 หน้า Install agent รูปที่ 4.23 Install agent success รูปที่ 4.24 หน้า Detecting CDH Versions รูปที่ 4.25 หน้า Inspect hosts for correctness รูปที่ 4.26 หน้า Select Service	42 43 44 45 46 46 47 48 49 50
รูปที่ 4.18 ทามแอก Repository รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials รูปที่ 4.22 หน้า Install agent รูปที่ 4.23 Install agent success รูปที่ 4.24 หน้า Detecting CDH Versions รูปที่ 4.25 หน้า Inspect hosts for correctness รูปที่ 4.26 หน้า Select Service รูปที่ 4.27 หน้า assign roles	42 43 44 45 46 46 47 48 49 50 51
รูปที่ 4.18 ทามเขา Repository รูปที่ 4.19 หน้า Accept JDk License รูปที่ 4.20 หน้า Single user mode รูปที่ 4.21 หน้า Enter Login Credentials รูปที่ 4.22 หน้า Install agent รูปที่ 4.23 Install agent success รูปที่ 4.24 หน้า Detecting CDH Versions รูปที่ 4.25 หน้า Inspect hosts for correctness รูปที่ 4.26 หน้า Select Service รูปที่ 4.27 หน้า assign roles	42 43 44 45 46 46 47 48 49 50 51 52
รูปที่ 4.18 หน้า Accept JDk License	42 43 44 45 46 46 47 48 49 50 51 52 53
รูปที่ 4.18 ทน้า Accept JDk License	42 43 44 45 46 46 47 48 49 50 51 52 51 52 53 54
รูปที่ 4.19 หน้า Accept JDk License. รูปที่ 4.20 หน้า Single user mode. รูปที่ 4.21 หน้า Enter Login Credentials. รูปที่ 4.22 หน้า Install agent. รูปที่ 4.23 Install agent success. รูปที่ 4.24 หน้า Detecting CDH Versions. รูปที่ 4.25 หน้า Inspect hosts for correctness. รูปที่ 4.26 หน้า Select Service. รูปที่ 4.27 หน้า assign roles. รูปที่ 4.28 หน้า Setup Database. รูปที่ 4.30 Test connection success. รูปที่ 4.31 หน้า Review Changes.	42 43 44 45 46 46 47 48 49 50 51 52 53 54 55

10

รูปที่ 4.33	Run service success
รูปที่ 4.34	Service are installed
รูปที่ 4.35	Administration_Settings
รูปที่ 4.36	Custom Service Description
รูปที่ 4.37	Path of parcel
รูปที่ 4.38	Host_Parcels
รูปที่ 4.39	Parcel_distribute
รูปที่ 4.40	Parcels_activate
รูปที่ 4.41	Add service
รูปที่ 4.42	เลือก Service StreamSets64
รูปที่ 4.43	เลือก Assign role for StreamSets
รูปที่ 4.44	Select hosts
รูปที่ 4.45	Cloudera start service StreamSets
รูปที่ 4.46	ติดตั้ง StreamSets สำเร็จ
รูปที่ 4.47	Hue login
รูปที่ 4.48	Home_Hue69
รูปที่ 4.48 รูปที่ 4.49	Home_Hue
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50	Home_Hue
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51	Home_Hue
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.52	Home_Hue
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.52 รูปที่ 4.53	Home_Hue
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.52 รูปที่ 4.53 รูปที่ 4.53	Home_Hue. .69 ใอกอนเมนู 3 ขีดของ hue. .69 Menu_File. .70 หน้าเก็บ File hdfs. .70 สร้าง File หรือ Directory. .71 หน้าเลือก File เพื่ออัพโหลด. .72
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.53 รูปที่ 4.53 รูปที่ 4.54 รูปที่ 4.55	Home_Hue. .69 ใอกอนเมนู 3 บีดของ hue. .69 Menu_File. .70 หน้าเก็บ File hdfs. .70 สร้าง File หรือ Directory. .71 หน้าเลือก File เพื่ออัพโหลด. .72 กลิกขวาเพื่อตั้งค่าไฟล์. .72
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.53 รูปที่ 4.54 รูปที่ 4.55 รูปที่ 4.55	Home_Hue. .69 ใอกอนเมนู 3 บีดของ hue. .69 Menu_File. .70 หน้าเก็บ File hdfs. .70 สร้าง File หรือ Directory. .71 หน้าเลือก File เพื่ออัพโหลด. .72 คลิกขวาเพื่อตั้งค่า ไฟล์. .72 เลือก replication. .73
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.53 รูปที่ 4.54 รูปที่ 4.55 รูปที่ 4.55 รูปที่ 4.56 รูปที่ 4.57	Home_Hue.
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.52 รูปที่ 4.54 รูปที่ 4.55 รูปที่ 4.55 รูปที่ 4.57 รูปที่ 4.57	Home_Hue.
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.52 รูปที่ 4.53 รูปที่ 4.54 รูปที่ 4.55 รูปที่ 4.55 รูปที่ 4.57 รูปที่ 4.58	Home_Hue
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.52 รูปที่ 4.53 รูปที่ 4.54 รูปที่ 4.55 รูปที่ 4.55 รูปที่ 4.58 รูปที่ 4.59 รูปที่ 4.60	Home_Hue.
รูปที่ 4.48 รูปที่ 4.49 รูปที่ 4.50 รูปที่ 4.51 รูปที่ 4.52 รูปที่ 4.53 รูปที่ 4.54 รูปที่ 4.55 รูปที่ 4.55 รูปที่ 4.59 รูปที่ 4.59 รูปที่ 4.60 รูปที่ 4.61	Home_Hue

iC.

รูปที่ 4.63	ตั้งค่าที่อยู่เมื่อเขียน File เสร็จ
รูปที่ 4.64	ข้อมูลทดสอบในFile CSV78
รูปที่ 4.65	เลือก Origin directory
รูปที่ 4.66	Data format Delimited
รูปที่ 4.67	Processors
รูปที่ 4.68	Field Type Converter
รูปที่ 4.69	Conversions
รูปที่ 4.70	เลือก Field to Convert
รูปที่ 4.71	Hive Metadata
รูปที่ 4.72	Hue Tables
รูปที่ 4.73	Table Databases
รูปที่ 4.74	Create a new database
รูปที่ 4.75	สร้าง database เรียบร้อย
รูปที่ 4.76	ตั้งก่า Hive metadata
รูปที่ 4.77	Stage destination
รูปที่ 4.78	Destination HDFS และ Hive Metastore85
รูปที่ 4.79	Hadoop fs general
รูปที่ 4.80	Hadoop fs หัวข้อ Output files
รูปที่ 4.81	Hive Metastore หัวข้อ hive
รูปที่ 4.82	การเข้า Impala
รูปที่ 4.83	Hue Impala
รูปที่ 4.84	ตัวอย่างการใช้ <mark>code</mark> SQL
รูปที่ 4.85	เข้า Ozzie
รูปที่ 4.86	Ozzie เปลี่ยน Document เป็น actions
รูปที่ 4.87	การเลือก Action ที่ต้องการทำงาน
รูปที่ 4.88	Pipeline id
รูปที่ 4.89	Script การสั่งstart StreamSets
รูปที่ 4.90	Ozzie เมื่อนำshellมาว่าง
ราใที่ 4 91	บันทึก flow oozie
а́ П I I.7 I	
	รูปที่ 4.63 รูปที่ 4.64 รูปที่ 4.65 รูปที่ 4.67 รูปที่ 4.67 รูปที่ 4.69 รูปที่ 4.70 รูปที่ 4.70 รูปที่ 4.71 รูปที่ 4.72 รูปที่ 4.73 รูปที่ 4.73 รูปที่ 4.74 รูปที่ 4.75 รูปที่ 4.75 รูปที่ 4.75 รูปที่ 4.76 รูปที่ 4.77 รูปที่ 4.78 รูปที่ 4.80 รูปที่ 4.81 รูปที่ 4.81 รูปที่ 4.81 รูปที่ 4.81 รูปที่ 4.81 รูปที่ 4.82 รูปที่ 4.83 รูปที่ 4.84 รูปที่ 4.84 รูปที่ 4.85 รูปที่ 4.85 รูปที่ 4.87 รูปที่ 4.84 รูปที่ 4.84 รูปที่ 4.83 รูปที่ 4.84 รูปที่ 4.84 รูปที่ 4.85 รูปที่ 4.87 รูปที่ 4.87 รูปที่ 4.88 รูปที่ 4.89 รูปที่ 4.90

1C

รูปที่ 4.93 เลือก schedule
รูปที่ 4.94 ตั้งก่า schedule97
รูปที่ ก.1 การติดตั้ง VMware (1)103
รูปที่ ก.2 การติดตั้ง VMware (2)104
รูปที่ ก.3 การติดตั้ง VMware (3)104
รูปที่ ก.4 การติดตั้ง VMware (4)105
รูปที่ ก.ร การติดตั้ง VMware (5)105
รูปที่ ก.6 การติดตั้ง VMware (7)106
รูปที่ ก.7 การติดตั้ง VMware (8)106
รูปที่ ก.8 การติดตั้ง VMware (9)107
รูปที่ ข.1 การติดตั้ง Linux (1)109
รูปที่ ข.2 การติดตั้ง Linux (2)109
รูปที่ ข.3 การติดตั้ง Linux (3)110
รูปที่ ข.4 การติดตั้ง Linux (4)
รูปที่ ข.5 การติดตั้ง Linux (5) 111
รูปที่ ข.6 การติดตั้ง Linux (6)
รูปที่ ข.7 การติดตั้ง Linux (7)112
รูปที่ ข.8 การติดตั้ง Linux (8)112
รูปที่ ข.9 การติดตั้ง Linux (9)113
รูปที่ ข.10 การติดตั้ง Linux (10) 113

STITUTE O





1.1 ชื่อและที่ตั้งของสถานประกอบการ

TC

ชื่อ	: บริษัท เอ-โฮสต์ จำกัด	
ที่ตั้ง	: 979/52-55 อาการ เอส เอ็ม ทาวเวอร์ ชั้น 21 ถนนพหล	โยธิน
	แขวงสามเสนใน เขตพญาไท กรุงเทพฯ 10400	
ติดต่อ	: 0-2298-0625-32 ต่อ 415 แฟกซ์ : 0-2298-005	53
เว็บไซต์	: http://www.a-host.co.th	



<mark>รูปที่ 1.1 แผน</mark>ที่ตั้ง บริษั<mark>ท เอ</mark>–โฮสต์ <mark>จำกั</mark>ด

1.2 ลักษณะธุรกิจของสถานประกอบการ หรือการให้บริการหลักขององค์กร

บริษัทเอ-โฮสต์ก่อตั้งบริษัทเมื่อปี 2542 ซึ่งเป็นบริษัทหนึ่งในเครือของบริษัท เมโทร ซิส เต็มส์ คอร์เปอเรชั่น (มหาชน) จำกัด และเป็นผู้เชี่ยวชาญด้านบริการจัดวางระบบไอที และบริการ เสริมต่าง ๆ สำหรับลูกก้ำตั้งแต่ธุรกิจขนาดย่อมไปจนถึงขนาดกลาง

ธุรกิจหลักของบริษัท เอ-โฮสต์ คือ การให้บริการโฮสติ้ง และบริการระบบไอทีด้วย ผลิตภัณฑ์ของออราเคิล (Oracle) และไอบีเอ็ม (IBM) ซึ่งเป็นซอฟท์แวร์สำหรับการวางแผนบริหาร ทรัพยากรขององค์กร (ERP) ระดับแนวหน้าของโลก

เอ-โฮสต์ถือกำเนิดขึ้นจากกลุ่มผู้เชี่ยวชาญด้านไอทีท่ามกลางภาวะเศรษฐกิจตกต่ำทั่ว ภูมิภาคเอเชียแต่ เอ-โฮสต์ ก็สามารถเติบโตได้อย่างรวดเร็ว และมั่นคงตั้งแต่แรกก่อตั้ง ด้วยจุดแข็ง ในฐานะผู้บุกเบิกธุรกิจโฮสติงเซอร์วิส พร้อมทั้งนำธุรกิจแนวใหม่อย่างการให้บริการแอพพลิเคชัน หรือ ASP (Application Services Providing) เข้ามาให้บริการแก่องค์กรธุรกิจเป็นรายแรกใน เมืองไทย

ธุรกิจการให้บริการแอพพลิเคชัน ในรูปแบบ ASP ของเอ-โฮสต์ไม่เพียงแต่ให้บริการค้าน แอพพลิเคชั่นด้านการดำเนินธุรกิจทางอิเล็กทรอนิกส์ระดับโลกของออราเคิลพร้อมโครงสร้าง พื้นฐานทางเทคโนโลยีสารสนเทศเท่านั้น แต่ยังมีบริการที่ครอบคลุมตั้งแต่การให้คำปรึกษา การ สนับสนุนและการให้บริการทั่วไปอย่างพรั่งพร้อมครบครัน รวมทั้งยังมีความยืดหยุ่นสูงมีการ ปรับเปลี่ยนบริการและทรัพยากรให้เมาะสมกับความต้องการ และสภาพงานที่แตกต่างกันของลูกค้า ในรายได้ระดับต่าง ๆ

76

ในการดำเนินธุรกิจของเอ-โฮสต์ตลอดระยะเวลา 10 ปี ไม่เพียงแต่ในฐานะผู้บุกเบิกธุรกิจ โฮสติ้งและธุรกิจการให้บริการแอพพลิเกชั่นในรูปแบบ ASP เท่านั้น แต่เอ-โฮสต์ยังได้ทำการติดตั้ง ระบบไอที รวมทั้งผลิตภัณฑ์ของออราเกิลให้กับลูกก้างนประสบความสำเร็จมาแล้วเป็นจำนวนมาก ซึ่งหลายรายเป็นหนึ่งในร้อยบริษัทชั้นนำของประเทศไทย แต่ที่สำคัญกว่านั้นก็คือการที่ เอ-โฮสต์ ได้สานสัมพันธ์กับลูกก้า และพันธมิตรทางธุรกิจอย่างแนบแน่นจนกลายเป็นหุ้นส่วนทางกลยุทธ์ และเป็นผู้สนับสนุนสำคัญที่มีส่วนช่วยผลักดันให้ธุรกิจจองลูกก้าเติบโตลู่กวามสำเร็จ

ปัจจุบันเอ-โฮสต์เป็นหนึ่งในบริษัทลูกของบริษัท คราก้อนวัน จำกัด ซึ่งเป็นบริษัทมหาชน ที่อยู่ในตลาดหลักทรัพย์แห่งประเทศไทย

เป็นเวลากว่า 10 ปีที่เอ-โฮสต์ และออราเคิลได้ดำเนินธุรกิจร่วมกันอย่างใกล้ชิด และถือเป็น พันธมิตรทางธุรกิจที่มีความแนบแน่นกันมานับตั้งแต่ก่อตั้งบริษัทปี 2542 จวบจนกระทั่งในปัจจุบัน ในปี 2011-HOST ได้ก้าวไปข้างหน้าเพื่อความท้าทายทางธุรกิจใหม่ที่จะเป็น "พันธมิตร ทางธุรกิจของไอบีเอ็มพรีเมียร์ (IBM Premier Business Partner)" เป็นตัวแทนจำหน่ายฮาร์ดแวร์ ของไอบีเอ็มและผลิตภัณฑ์ซอฟต์แวร์ ที่จะสามารถเสริมสร้างผลิตภัณฑ์และผลงานบริการของเรา เพื่อให้ลูกค้าได้รับเทคโนโลยีที่ดีที่สุดอยู่ตลอดเวลาและช่วยส่งเสริมการเจริญเติบโตของลูกค้าอย่าง รวดเร็วและมั่นคง

ด้วยความมุ่งมั่นในการนำเสนอผลิตภัณฑ์ของออราเคิล และ ไอบีเอ็ม ผ่านการให้บริการ แอพพลิเคชันในรูปแบบของ ASP ในฐานะที่เอ-โฮสต์เป็นผู้บุกเบิกธุรกิจคังกล่าว และเพิ่มศักยภาพ ในการคำเนินธุรกิจของลูกค้าได้อย่างเต็มประสิทธิภาพ และเหมาะสม ทำให้ได้รับรางวัลแห่ง ความสำเร็จและได้รับการยกย่องมาอย่างต่อเนื่อง



(

รูปที่ 1.2 A-HOST Proud Awards

1.3 รูปแบบการจัดองค์กรและการบริหารองค์กร





คุณบุญประสิทธิ์ ตั้งชัยสุข กรรมการพู้จัดการ



คุณเลิศ รักษ์ศิริวณีช กรรมการพู้จัดการ ABCs Company Limited

T





คุณวิชัย วงศ์จริยกุล พู้อำนวยการฟ่ายให้คำปรึกษา

รูปที่ 1.3 คณะผู้บริหารบริษัท เอ-โฮสต์ จำกัด แต่ละแผนก



คุณประสงค์ เอื้อสุริยนันท์ ยการพ่าย Hosting and Outs งอ่านวยก



คุณกนกวรรณ หะลีห์รัตนวัฒนา พู้อำนวยการฟ่ายการตลาด



4

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

10

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

1.7 ที่มาและความสำคัญของปัญหา

ในปัจจุบันหลาย<mark>องก์</mark>กรในประเทศไทยและทั่วโลก ก<mark>ำลังก</mark>ล่าวถึงข้อมูลขนาดใหญ่ Big Data เพราะองก์กรต่างๆทั้งในอดีตและปัจจุบัน ข้อมูลนั้นถือว่าเป็นสิ่งที่สำคัญในการบอกถึงสิ่งที่ กำลังเกิดขึ้น บริษัท A-HOST จำกัด จึงเริ่มทำ Big Data โดยด้องการระบบ Big Data และเครื่องมือ ต่าง ๆ ของ Big Data และวิธีการใช้งาน

1.8 วัตถุประสงค์หรือจุดมุ่งหมายของโครงงาน

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

1.9 ผลที่คาดว่าจะได้รับจากการปฏิบัติงานหรือโครงงานที่ได้รับมอบหมาย

- นักสึกษานำความรู้ที่ได้จากการปฏิบัติงานสหกิจไปใช้ประกอบอาชีพในอนาคต
- 2. นักศึกษาสามารถทำงานร่วมกันผู้อื่นได้เป็นอย่างดี
- 3. นักศึกษามีความรู้และทักษะเฉพาะทางในสายงานนี้มากยิ่งขึ้น
- 4. นักสึกษามีความรับผิดชอบในหน้าที่การงานที่ได้รับมอบหมาย
- 5. ระบบสามารถใช้งานได้จริง
- 6. ระบบสามารถโอนย้ายข้อมูลตามความต้องการของลูกค้าได้
- 7. ระบบมีการโอนย้ายข้อมูลได้อย่างครบถ้วน

1.10นิยามศัพท์เฉพาะ

1. BI (Business Intelligence) หมายถึง ชุดของแนวกิดและกระบวนทัศน์ที่จะพัฒนา กระบวนการ ตัดสินใจของธุรกิจโดยอาศัยข้อมูลที่เป็นข้อเท็จจริงจากฐานข้อมูล

2. Bigdata หมายถึง ปริมาณข้อมูลจำนวนมหาศาลที่มีอยู่ในบริษัทของคุณทุกรูปแบบ ไม่ว่า แหล่งที่มาจะมาจากภายในบริษัทหรือภายนอกก็ตาม ทั้งนี้แบ่งออกเป็นข้อมูลที่มีโครงสร้างชัดเจน (Structured Data) และข้อ<mark>มูลที่</mark>มีโครงสร้างไม่ชัดเจน (Unstructured Data)

3. StreamSets คือ <mark>เครื่</mark>องมือสำหรั<mark>บ</mark>การดึงข้อมู<mark>ล แ</mark>ปลงข้<mark>อมูล</mark>และ โหลดข้อมูล

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

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

บทที่ 2 ทฤษฎีและเทคโนโลยีที่ใช้ในการปฏิบัติงาน

ในการปฏิบัติงานสหกิจศึกษาครั้งนี้ เป็นการนำความรู้ทางด้านทฤษฎีและเทคโนโลยีมาใช้ ในการปฏิบัติงานทุกส่วนตลอดการปฏิบัติงานสหกิจศึกษา ซึ่งเป็นการนำความรู้ทั้งที่เลยเรียนมา ประยุกต์ใช้และเป็นการศึกษาเรียนรู้สิ่งใหม่ ๆ ที่ได้จากการปฏิบัติงาน

2.1 ระบบปัญญาธุรกิจ (Business Intelligence: BI)

10

BI คือ การนำเสนอข้อมูลเพื่อช่วยในการวางแผนการตัดสินใจหรือตอบกำถามเชิงธุรกิจ ให้กับผู้บริหาร ทำให้องก์กรสามารถกาดการณ์หรือพยากรณ์ความต้องการของผู้บริโภคได้อย่าง ถูกต้องแม่นยำ ส่งผลให้ประสิทธิภาพการทำงานขององก์กรสูงขึ้น

Business Intelligence หรือที่เรียกว่า "BI" นั้นได้ถูกคิดค้นโดย Howard Dresner ในช่วงปี 1990 Howard Dresner ให้ความหมาย คำว่า "ระบบปัญญาธุรกิจ" (Business Intelligence) หรือ BI หมายถึง "ชุดของแนวคิดและกระบวนทัศน์ที่จะพัฒนากระบวนการตัดสินใจของธุรกิจโดยอาศัย ข้อมูลที่เป็นข้อเท็จจริงจากฐานข้อมูล"



2.1.1 ลักษณะสำคัญของระบบ BI คือ

ระบบงานของ BI

จุดเด่นของ BI คือ ใช้งานง่ายผู้ใช้ไม่ต้องมีความรู้ด้านฐานข้อมูล ก็สามารถใช้งานได้ เพียงเลือก ข้อมูลที่ต้องการก็สามารถได้ผล<mark>ดัพธ์ตามต้องการ ข้อมู</mark>ลมีความถูกต้องแม่นยำทำให้สามารถใช้ ข้อมูลเพื่อช่วยในการตัดสินใจได้รวดเร็วกว่ากู่แข่ง ทั้งในเชิงกว้าง และเชิงลึก สามารถดึงข้อมูลจาก ฐานข้อมูลที่หลากหลายมาทำการวิเคราะห์โดยไม่มีการเขียนโปรแกรม



รูปที่ 2.2 ระบบงานของ BI

2.1.2 กระบวนการทำงานของ BI

ก่อนที่จะเกิดเป็น BI ได้ต้องมีกระบวนการทำงานต่าง ๆ ที่ทำงานร่วมกันเพื่อให้ข้อมูล ออกมาตามที่ต้องการและวางแผนไว้ล่วงหน้าได้โดนต้องมีกระบวนการต่าง ๆ ดังต่อไปนี้

1) การกำหนดแหล่งข้อมูล (Data Sources) แบ่งออกเป็น 2 ประเภท คือ

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

- แหล่งข้อมูลภา<mark>ยนอ</mark>ก (External Data Sources) ใด้แก่ ข้อมูลสถิติจากสถาบันต่าง ๆ ข้อมูล ของโครงการสารสนเทศอื่<mark>น ๆ</mark> บทวิเคราะห์และบทความวิชาการ<mark>ต่าง</mark> ๆ

2) การออกแบบคลังข้อมูล (Data Warehouse Design)

 เป็นที่จัดเก็บข้อมูลที่นำมาจากแหล่งข้อมูลภายในองค์กรจากฐานข้อมูลการใช้งาน ประจำวันหรือฐานข้อมูลปฏิบัติการ (Operational Database) หรือมาจากฐานข้อมูลภายนอกองค์กร (External Database)

ข้อมูลในคลังข้อมูลจะถูกนำมาใช้เพื่อสนับสนุนการตัดสินใจบริหารงานของผู้บริหาร



- เช่น ระบบสนับสนุนการตัดสินใจ (Decision Support System) หรือ ระบบการจัดการ สารสนเทศ (Management Information System)

รูปที่ 2.3 Data Warehouse

รูปแบบการนำข้อมูลมาประมวลผลแบ่งออกเป็น 2 ประเภทใหญ่ ๆ ดังนี้ OLTP (Online Transaction Processing) คือ การประมวลผลข้อมูลตามลักษณะการ ปฏิบัติงานประจำวันของหน่วยงานนั้นจากฐานข้อมูลเช่น ERP (Enterprise resource planning), POS (point of sale)

10

OLAP (Online Analysis Processing) เป็นเทค โนโลยี ที่ใช้คึงข้อมูลจาก Data Warehouse นำไปวิเคราะห์เพื่อใช้สนับสนุนการตัดสินใจของผู้บริหารจากคลังข้อมูลรูปแบบในการวิเคราะห์มี ดังนี้

- BI (Business In<mark>tellig</mark>ence) ซอฟ<mark>ต์แวร์ที่นำข้อมูล</mark>ที่มีอยู่<mark>เพื่อ</mark>จัดทำรายงานในรูปแบบต่าง ๆ ที่เหมาะสมกับมุมมองในการวิเกราะห์และตรงตามกวามต้องการของผู้ใช้งานตามแต่ละแผนก

- Data mining ระ<mark>บบช่</mark>วยดูแนวโน้มในอนาก<mark>ตุกว</mark>ามสัมพั<mark>นธ์</mark>ของข้อมูล

- DSS (Decision Support System) ระบบสนับสนุนผู้บริหารเพื่อช่วยผู้บริหารในการ ตัดสินใจเชิงกลยุทธ์

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 วิธีการนี้จะช้ากว่า แบบ MOLAP แต่ข้อมูลที่ได้ทันสมัยเสมอ

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

- HOLAP (Hybrid OLAP) คือ OLAP หรือ Cube ที่มีการแบ่งพื้นที่ออกเป็นส่วน ๆ โดยแต่ ละส่วนอาจใช้วิธีการจัดเก<mark>็บข้อ</mark>มูลแบบ MOLAP และบางส่วนก็จั<mark>ดเก็บ</mark>แบบ ROLAP Design Data Warehouse ซึ่งการออกแบบค<mark>ลังข้อมูลมีอยู่ 2</mark> 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 จะเป็นศูนย์ รวมข้อมูลเพียงTable เดียวและจำนวนมุมมองจะเท่ากับจำนวนของ Dimension ที่รายล้อมอยู่ ซึ่ง ลักษณะแบบนี้ จะช่วยเพิ่มความสามารถในการ Query ข้อมูลได้ง่ายและรวดเร็ว



รูปที่ 2.5 กลังข้อมูลแบบ Star Schema หรือ Multidimensional Schema

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





องค์ประกอบสำค<mark>ัญใน</mark>ตารางมี 2 <mark>ป</mark>ระเภท <mark>คือ</mark>

- Fact Table เป็น<mark>ตารา</mark>งหลัก ซึ่งม<mark>ีลั</mark>กษณะคล้<mark>ายกับตารางปร</mark>ะเภท Transaction ของ OLTP โดยภายในจะประกอบด้วยคอลัมน์ที่สำคัญ 2 ประเภทคือ

- Key เป็นกอลัมน์ที่ใช้เชื่อมโยงไปยัง Dimension Table ต่าง ๆ ดังนั้นจำนวนกอลัมน์ของ Fact Table Key จะเพิ่มขึ้นตามจำนวนของ Dimension Table และกอลัมน์ทั้งหมดนี้สามารถ นำไปใช้สร้างให้เป็น Primary Key ได้อีกด้วย

- Measure เป็นข้อมูลตัวเลข ทำหน้าที่เก็บจำนวน หรือปริมาณที่เกิดขึ้นของแต่ละ Transaction นอกจากนี้ยังเก็บผลลัพธ์ที่ได้จากการคำนวณ Dimension Tables เป็นตารางข้อมูลที่เก็บการอธิบาย Entity ต่าง ๆ โดยประกอบด้วยคอลัมน์ที่เป็น Key เพื่อเชื่อมโยงไป Fact Table Key และคอลัมน์ที่ให้ความหมายเพิ่มเติมแก่ Entity สามารถนำไป สร้างเป็น Dimension ของ OLAP Cube ตามลักษณะต่าง ๆ ดังนี้

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

- Parent-Child Dimension มีลักษณะคล้ายกับ Standard Dimension แต่ภายในจะมี ความสัมพันธ์ระหว่างภายในกันเอง

Data Mart คือ คลังข้อมูลขนาคเล็กมีการเก็บข้อมูลที่มีลักษณะเฉพาะเจาะจง เช่น เก็บข้อมูลส่วน ของการเงิน ส่วนของสินค้าลงคลัง ส่วนของการขาย เป็นต้น ซึ่งทำให้การจัดการข้อมูลการนำเอา ข้อมูล ไปสร้างความสัมพันธ์และวิเคราะห์ต่อ ก็ง่ายขึ้น เมื่อมีข้อมูลใน Data Warehouse แล้ว สามารถสร้าง Data Mart ขึ้นมาได้ ซึ่ง Data Mart นั้นถูกออกแบบมาเพื่อช่วยเหลือผู้ใช้ในการตอบ คำถามทางธุรกิจตามที่ผู้ใช้ต้องการ ขั้นตอนการย้ายข้อมูลจาก Data Warehouse เข้าสู่ Data Mart เรียกว่า Data Deliver



รูปที่ 2.7 Data Mart

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 การนำข้อมูลมาจัครูปแบบให้ถูกต้องสอดคล้องกัน
- Data Mapping การทำให้ข้อมูลให้อยู่ในรูปแบบเดียวกัน

- Data Cleansing การตรวจสอบและแก้ไขข้อมูลให้ถูกต้อง

- Load คือ การนำข้อมูลที่ผ่านการ Transform แถ้ว เข้าสู่ Data Warehouse โดยทั่วไปแถ้ว ETL Tools สามารถทำได้ดังนี้

- Data Cleansing-ตรวจสอบความถูกต้องของข้อมูล รวมทั้งกำจัดข้อมูลที่ผิดพลาด

- Data Transformation–การแปลงข้อมูล หรือจัดรูปแบบข้อมูลเพื่อให้สามารถนำไป วิเคราะห์ได้ง่ายขึ้น

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



รูปที่ 2.8 ETL Extract-Transform-Load

4) การทำข้อมูลให้อยู่ในรูปแบบ Multidimensional Model หรือ Cube รูปแบบการทำให้ข้อมูลเกิดมิติขึ้นในหลาย ๆ ด้านก่อนจะนำไปสร้างเป็นรายงานใน รูปแบบต่าง ๆ โดยอาศัยเกรื่องมือที่ช่วยในการ Query ข้อมูลที่ต้องการ เช่น Oracle Essbase, Cognos Transformer





5) การออกแบบและสร้างรายงาน (Interactive Report)

1G

รายงานที่นำเสนอมักจะเป็นผลการคำเนินงานตามตัวบ่งชี้การคำเนินงานต่าง ๆ ของ หน่วยงานหรือการติดตามก่าเป้าหมายของการดำเนินงาน ที่สำคัญการนำเสนอรายงานมักจะอยู่ใน รูปแบบของกราฟเพื่อทำให้เกิดความเข้าใจได้ง่ายผ่าน Dashboard ที่ผู้ใช้สามารถเข้าถึงผ่านหน้า เว็บไซต์ที่จัดทำไว้ เช่น OBI Report, Cognos Report เป็นต้น



รูปที่ 2.10 กระบวนการทำงานของ BI AN INSTITUTE OF

2.1.3 ประโยชน์ของ BI

10

 ช่วยเพิ่มศักยภาพในการตัดสินใจให้ถูกต้องและรวดเร็วจากข้อมูลที่มีอยู่ โดยเห็น ภาพพจน์ของข้อมูลที่มีก่อนการตัดสินใจ

เพิ่มประสิทธิภาพในการแลกเปลี่ยนข้อมูลภายในองค์กร โดยสามารถแลกเปลี่ยนข้อมูล
 ภายในผ่านเครือข่ายได้ในแบบอัตโนมัติ

ลดต้นทุนทั้งด้านเงินและเวลาในการเข้าถึงข้อมูลองก์กร

 ช่วยให้ผู้ใช้สามารถตอบคำถามทางธุรกิจที่สำคัญ และช่วยให้สามารถ รวบรวมและปรับ ข้อมูลตามต้องการเพื่อสร้างแนวกิดที่แตกต่าง

- สำรวจข้อมูลทุกชนิดจากทุกแง่มุม

- วิเคราะห์ข้อเท็จจริงและคาดการณ์นัยแฝงเชิงยุทธวิธีและเชิงกลยุทธ์

15

2.2 ข้อมูลขนาดใหญ่ (Bigdata)

Bigdata คือ ปริมาณข้อมูลที่มีขนาดใหญ่มหาศาลเกินกว่าขีดความสามารถในการ ประมวลผลของระบบฐานข้อมูลธรรมดาที่จะรองรับได้ปริมาณข้อมูลที่มีขนาดใหญ่มาก ๆ จะมี อัตราการเพิ่มข้อมูลได้อย่างรวดเร็วมากและจะมีรูปแบบที่มีโครงสร้างหรือไม่มีโครงสร้าง ซึ่งไม่ สามารถอยู่ในระบบฐานข้อมูลที่จะจัดเก็บข้อมูลที่ได้ โดย Bigdata นั้นมีลักษณะพื้นฐานอยู่ 3 ลักษณะหลัก ๆ อธิบายโดยใช้ 3V Model ได้แก่ ปริมาณ (Volume) ความเร็ว (Velocity) และความ หลากหลาย (Variety)

(ula a)



รูปที่ 2.11 3V Model of Big data

ปริมาณ (Volume<mark>) คือ</mark>ปริมาณของข้อมูลที่<mark>มีจำนว</mark>นมหาศาล ซึ่งในอดีตการเก็บข้อมูลที่มี ขนาดใหญ่นั้นมักจะเป็นปัญหา แต่ในปัจจุบันเทคโนโลยี Big Data ทำให้การจัดการทำข้อมูลที่มี ขนาดใหญ่นี้เป็นเรื่องที่ง่ายขึ้นกว่าเดิมมาก

ความเร็ว (velocity) คือการเข้าถึงข้อมูลและจัดการข้อมูลนั้นจะต้องทำได้แบบทันทีทันใด ข้อมูลที่เกิดขึ้นมากวรถูกเก็บและวิเกราะห์ให้ตรงเวลา

ความหลากหลาย (Variety) คือรูปแบบของข้อมูลที่แตกต่างกันออกไป ไม่ว่าจะเป็น ตัวอักษร วิดีโอ รูปภาพ หรือกี่คือข้อมูลแบบมีโกรงสร้าง 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) เป็น โปรแกรมฟรีแวร์สำหรับจำลอง ระบบคอมพิวเตอร์ เป็นซอฟต์แวร์สำหรับใช้ทำการจำลองระบบคอมพิวเตอร์ (Virtualization) บน ระบบ x86 และ AMD64/Intel64 ลักษณะเดียวกับโปรแกรม VMware Workstation (เป็นโปรแกรม เชิงพาณิชย์ต้องซื้อจึงจะใช้งานได้เต็มพังก์ชัน) 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 สามารถใช้งานได้ฟริโดยไม่มีค่าใช้จ่ายภายใต้ ใลเซนส์แบบ GNU General Public License (GPL) เป็นซอฟต์แวร์ที่มีประสิทธิภาพสูงรองรับการ ใช้งานได้ทั้งในเอนเทอร์ไพรส์ (Enterprise) และการใช้งานภายในบ้าน และยังมีฟีเจอร์ให้ใช้งาน หลากหลายและที่สำคัญเป็นโซลูชั่นระดับมืออาชีพที่ใช้งานได้ฟรี

Virtual Box คือโปรแกรมที่ใช้ในการจำลองเครื่องคอมพิวเตอร์ขึ้นมาอีกเครื่องหนึ่งโดย การแบ่งทรัพยากรจากระบบหลักไปใช้เช่น CPU, RAM, VGA, และ HDD โดยจุดมุ่งหมายหลักของ โปรแกรมนี้คือการติดตั้ง ระบบปฏิบัติการขึ้นมาอีกตัวหนึ่งเพื่อใช้งานที่แตกต่างกันไป

(

cloudera

รูปที่ 2.13 Logo Cloudera

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

- การประมวลผลข้อมูลภายในหน่วยความจำ : ด้วยTools Apache Spark

- การวิเคราะห์ที่รวดเร็วด้วย SQL: ค่าแฝงต่ำสุด เห็นพ้องที่สุดสำหรับโซลูชันการวิเคราะห์ ข้อมูลเพื่อเพิ่มประสิทธิผลทางธุรกิจด้วย Apache Impala

- การก้นหาแบบคั้งเดิม: การเข้าถึงข้อมูลของผู้ใช้ได้อย่างสมบูรณ์ซึ่งติดตั้งมาพร้อมกับ แพลตฟอร์ม Apache Solr

- อุปกรณ์เก็บข้อมูลเชิงวิเคราะห์ที่ปรับปรุงให้ทันสมัยอยู่เสมอ: อุปกรณ์จัดเก็บข้อมูล Hadoop ที่สามารถวิเคราะห์ข้อมูลที่เปลี่ยนแปลงได้อย่างรวดเร็วเมื่อใช้งานร่วมกับ Apache Kudu เท่านั้น

- การปรับแต่งข้อ<mark>มูลเชิ</mark>งรุกให้เกิด<mark>ประสิทธิผ</mark>ลสูง: Cloudera Navigator Optimizer (limited beta) ช่วยในการปรับแต่งข้อมูลและปริมาณงานเพื่อประสิทธิภาพการคำเนินงานสูงสุดในการ ทำงานร่วมกับ Hadoop

2.3.3 Apache Hadoop

10



รูปที่ 2.14 Logo Hadoop

Hadoop เป็นซอฟท์แวร์ประเภท open source ที่จัดทำขึ้นเพื่อเป็นแพลตฟอร์มในการจัดเก็บ ข้อมูล ซึ่งมีกรอบการทำงานเพื่อใช้ในการจัดเก็บข้อมูลและประมวลผลข้อมูลที่มีขนาดใหญ่ ที่เรา เรียกกันว่า Big Data ซึ่ง Hadoop ก็สามารถปรับขยาย ยืดหยุ่น เพื่อรองรับข้อมูลที่มีจำนวนมากมาย มหาศาลได้ ทั้งนี้ก็เพราะมันมีการกระบวนการประมวลผลที่แข็งแกร่งมากซึ่งเป็นผลมาจากการ ประมวลผลข้อมูลแบบกระจายผ่านเครื่องคอมพิวเตอร์ที่ถูกจัดอยู่ในรูปแบบ Cluster อันนำไปสู่ ความสามารถในการรองรับข้อมูลที่ไม่จำกัดแถมยังมีความน่าเชื่อถือสูงอีกด้วย

ประวัติความเป็นมาของ Hadoop ต้องข้อนกลับไปในปี 2006 หลังจากที่ World Wide Web เติบโตจนถึงจุดที่การใช้งานอินเตอร์เน็ตมีการขยายวงกว้างออกไปเรื่อยๆ ผู้ค้นค้นหาข้อมูลต่างๆ พอๆกับที่มีการป้อนคอนเท้นท์และข้อมูลเข้าไป ในปีนั้นเองที่ Google เริ่มมีการทำงานเกี่ยวกับการ จัดเก็บข้อมูลและการประมวลผลข้อมูล Yahoo และทีมผู้พัฒนาซอฟท์แวร์จึงได้มีการเริ่มด้นพัฒนา Hadoop ขึ้น ซึ่งชื่อนี้มีที่มาจากชื่อของเล่นของลูกชายหัวหน้าทีมผู้พัฒนานั้นเอง จากนั้นในปี 2008, Yahoo ก็ได้ปล่อย Hadoop ออกสู่สาธารณชนในฐานะ open-source project ต่อมา Hadoop จึงตกอยู่ ภายใต้การดูแลขององค์กรที่ไม่แสวงหาผลกำไรอย่าง Apache Software Foundation (ASF) อย่างที่ เห็นในปัจจุบัน

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



รูปที่ 2.15 หน้าโปรแกรม HUE

Hue หรือ Hadoop User Experience

10

เป็นหน้าจอ Web UI ของค่าย Cloudera ทำให้สามารถจัดการ Application ตัวอื่นๆ เช่น Hive,impala,pig,sqoop,Hbase,HDFS เป็นค้น ผ่านทาง Web Browser ได้อย่างสะดวกและง่ายกว่าใช้ กำสั่ง Command Line 2.3.5 Apache Hive



รูปที่ **2.16** Logo HIVE

Hive

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

2.3.6 โปรแกรม Impala

10



รูปที่ 2.17 Logo Impala

Impala

เป็นเครื่องมือที่ทางค่าย Cloudera ทำการ Build เข้ามาในตัว Cloudera Hadoop นั่นเอง โดย มีการทำงานคล้ายๆกับ Hive แต่ที่แตกต่างกันคือ Impala จะทำงานกับข้อมูลที่อยู่บนMemory และ ติดต่อข้อมูล HDFS โดยตรงโดยที่ไม่ต้องผ่าน MapReduce ซึ่งจะทำให้Impala เร็วกว่าHive อย่าง แน่นอน

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

2.3.7 โปรแกรม Ozzie

รูปที่ 2.18 Logo OOZIE

Ozzie

10

เป็นเครื่องมือในการทำ Workflow จะช่วยให้เราเอาคำสั่งประมวลผลต่างๆของระบบ Hadoop เช่น Map/Reduce, Hive หรือ Pig มาเชื่อมต่อกันในรูปของ Workflow ได้
Best Che flag A distributed streaming platform

รูปที่ 2.19 Logo Kafka

Kafka

(0)

เป็น Streaming Platform ตัวหนึ่งทำหน้าที่เป็น Broker รับ Message จากที่นึงไปยังอีกที่ หนึ่ง 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

 ผู้ใช้จำนวน 300 million users events ทุกวันๆ ทำให้ในบางที่มันเกิดปัญหาเรื่องของ data lost(รวมไปถึงการทำ internal services messaging ด้วย นึกถึงสภาพบริษัทเรามี cluster A, cluster B, cluster C เยอะแยะไปหมด จะทำการติดต่ออย่างไรให้มีประสิทธิภาพ)

2. ปัญหาเรื่องของ การย่อยข้อมูลขนาดใหญ่ ขนาดนั้นเวลาเอามาใช้งานอีกด้วย

จึงเกิดโปรเจกนี้ขึ้นมาและถูกเผยแพร่ค่อ ผ่านทาง Apache Incubator ตั้งแต่ปี 2012 จากนั้น จึงได้แยกบริษัทออกมาจาก LinkedIn ก่อตั้งเป็น บริษัท Confluent เพื่อพัฒนา Kafka โดยเฉพาะ โดยชื่อ Kafka มาจากนักเขียนนาม Franz Kafka โดยเลือกชื่อ Kafka เพราะมันถูก optimize สำหรับ การเขียน เหมือนกับงานข<mark>อง F</mark>ranz Kafka

Kafka คือย่างไร

ยกตัวอย่าง ถ้ามีระบบสองตัวให้ติดต่อกันระหว่างสองตัวนี้ ก็จะติดต่อกันแบบ peer to peer แต่ถ้าหากมีมากขึ้นก็ต้องทำการเชื่อมต่อในแต่ละระบบมากขึ้น



รูปที่ 2.20 ตัวอย่างserverส่งข้อมูลเข้า Kafka

ดังนั้น Apache Kafka จึงถูกใช้เป็นตัวกลางในการเชื่อมต่อ เพื่อช่วยแยกการสื่อสารระหว่าง ระบบแต่ละตัว 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)

- มีประสิทธิภา<mark>พด้า</mark>นความเร็ว (latency น้อยกว่า 10m<mark>s)</mark>
- มีระบบที่ใหญ่ๆ ที่ใช้ Apache Kafka อยู่มาก เช่น Linkedin, Netflix, AirBnB, Yahoo, Wallmart หรือ LINE

สิ่งที่ Apache Kaf<mark>ka สา</mark>มารถทำได้

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

บทที่ 3 แผนงานการปฏิบัติงานและขั้นตอนการดำเนินงาน

3.1 แผนงานการฝึกงาน

ตารางที่ 3.1 ตารางแผนการฝึกงาน

หัวข้องาน	เรื่	ลือา	ามู่	1	เรื	้อเ	เพื่	2	เร	ลือเ	เทื่	3	เดิ	้อน	เทื่	4	
ศึกษา Oracle Business Intelligence Enterprise				y		/							1				
ศึกษา Concept Big Data							-	/	_	•							
Installation Program Linux สำหรับทดสอบ Big Data											1						
Install LibreOffice on Linux												Ì	2	\ . *			
แปลงข้อมูลผ่านเครื่องมือ LibreOffice โดยใช้														_			
Script แปลง													-				-
ศึกษา Python และ Modify Script														Ş			
ศึกษาเครื่องมือ ETL StreamSets																	
ศึกษาเครื่องมือทางด้าน Big Data ได้แก่																	
Hue, hive และImpala																	
ศึกษาเครื่องมือต่อย <mark>อ</mark> คเพิ่ม <mark>เติม</mark> จากโ <mark>คร</mark> งกา <mark>ร Big</mark>																	
Data ได้แก้ Ozzie,Kafka														(2		
ทำเอกสารเครื่องมือ Big Data													(
นำเสนอเครื่องมือ Big Da <mark>ta มา</mark> ปรับใช้ในธุ <mark>ร</mark> กิจ														/			
ติดตั้งโปรแกรม Cloudera ขึ้น Cloud A-Host									1		1	L)		<		
ทำเอกสารการติดตั้งCloudera																	1
Monitoring Report TAT			_	-			.<);		1							

3.2 รายละเอียดที่นักศึกษาปฏิบัติในการฝึกงาน

เนื่องจากผู้จัดทำโครงงานได้เข้าร่วมสหกิจในช่วงที่แผนก IBM Channel (BI) เริ่มต้นการ รับทำBig Data และได้มี Project Big Data ซึ่งอยู่ในช่วงImplement ดังนั้นผู้จัดทำโครงงานจึงได้รับ มอบหมายให้ทำการช่วยนำข้อมูลจากต้นทาง ไปปรับปรุงซึ่งเป็นกระบวนการ ETL ใช้โปรแกรม StreamSets และส่งไปยังปลายทางที่เก็บ คือ Hadoop distributed file system (HDFS) และ ทำการศึกษาเครื่องต่างๆ เพื่อเตรียมความพร้อมในการรับมือกับ requirement ของลูกค้า และต่อยอด ศึกษาเครื่องมือ Big Data เพิ่มเติมเพื่อรับมือกับงานของลูกค้าท่านอื่นต่อไป และทำการติดตั้ง Cloudera ซึ่งเป็นตัวจัดการให้เครื่องมือ Big Data ต่างๆ อยู่ในสภาพแวคล้อมเดียวกันและง่ายต่อการ จัดการ 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

ใด้ศึกษากระบวนการ ETL (Extract transform load) เพื่อที่จะสามารถนำข้อมูลจากต้นทาง ไปปรับปรุง และส่งไปยังปลาย ผู้จัดทำได้รับมอบหมายให้นำข้อมูลจากต้นทางไปเก็บใน HDFS ซึ่ง เป็นที่เก็บข้อมูลแบ<mark>บ</mark>กระจ<mark>าย ที่</mark>เป็นรูปแบ<mark>บการเก</mark>็บของ Big Data ด้วย</mark>การใช้โปรแกรม StreamSets

3.3.3 ศึกษาการใช้และทำ<mark>ความ</mark>เข้าใจ StreamSets

ศึกษาการใช้งาน StreamSets ซึ่ง<mark>เ</mark>ป็นเครื่องมือ ETL เพื่อย้ายข้อมูลของลูกค้า เนื่องจาก ผู้จัดทำโครงงานไม่เลยเรียนรู้มาก่อน จึงต้องเรียนรู้จากเอกสารและทำแบบฝึกหัดที่พี่พนักงานให้มา เพื่อไปทำการย้ายข้อมูลของลูกค้า

3.3.4 ศึกษาเครื่องมือและทำความเข้าใจ Hue

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

3.3.5 ศึกษาเครื่องมือและทำความเข้าใจ Hive,Impala

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

3.3.6 ศึกษาเครื่องมือ Ozzie

ศึกษาการใช้งาน Ozzie ซึ่งเป็นเครื่องมือสำหรับกำหนดการทำงานอัตโนมัติ เพื่อให้ สามารถกำหนดเวลาทำงานของเครื่องต่างๆ ได้

3.3.7 ศึกษาระบบ Kafka

(0)

ศึกษาระบบ Kafka เพื่อไม่ให้ข้อมูลขาดหายหรือมาไม่สมบูรณ์ระหว่างการรับส่งข้อมูล และนำไปต่อยอดการทำ Real-time processing กับ big data

บทที่ 4

สรุปผลการดำเนินงาน การวิเคราะห์และสรุปผลต่าง ๆ

4.1 ขั้นตอนและผลการดำเนินงาน

ในส่วนของขั้นตอนการคำเนินงานของBig Dataนั้นคือ การคึงข้อมูลจากแหล่งข้อมูลต้น ทาง สู่คลังข้อมูลปลายทางนั้นคือHadoop distributed file system (HDFS) ทำการศึกษาหาความรู้ที่ จะต้องนำไปใช้ประมาณ 2 สัปคาห์ หลังจากนั้นจึงเริ่มปฏิบัติงานจริง

โดยหลักแล้ว จะแบ่งการทำงานออกเป็น 6 ขั้นตอน คือ

4.1.1 การติดตั้งโปรแกรม Cloudera Manager บน Centos 7

ในส่วนนี้นั้นเป็นขั้นตอนการติดตั้งระบบ hadoop เพื่อให้ทำงานอย่างมีประสิทธิภาพจึงใช้ Cloudera Platform ซึ่งเป็นตัวทำให้เครื่องมือต่างๆ อยู่ในสภาพแวดล้อมเดียวกันและง่ายต่อการ จัดการ

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 และ โหลด เมื่อโหลด เสร็จเปลี่ยนไฟล์นามสกุล.sha1 เป็น .sha

2) CM5 repo-tarball

เป็นไฟล์หลักขอ<mark>ง Clo</mark>udera Manager สำหรับติ<mark>ด</mark>ตั้ง Clo<mark>uder</mark>a Agent คาวน์โหลดตามลิ้งนี้ https://archive.cloudera.com/cm5/repo-as-tarball/5.15.0/

3) Cloudera package

เป็น package ขอ<mark>ง Cloud</mark>era คาว<mark>น์โห</mark>ลดได้ตามลิ้งนี้

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-archivedownloads-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** การตั้งค่าเ<mark>บื้อง</mark>ต้น
- 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 เป็นตัวคักจับ package ชนิดหนึ่งของค่าย CentOS

ทำการเปลี่<mark>ยนดังรูปที่ 4.3</mark>

[root@localhost ~]# vi /etc/sysconfig/selinux

```
# This file controls the state of SELinux on the system.
# SELINUX= can take one of these three values:
# enforcing - SELinux security policy is enforced.
# permissive - SELinux prints warnings instead of enforcing.
# disabled - No SELinux policy is loaded.
SELINUX=disabled
# SELINUXTYPE= can take one of three two values:
# targeted - Targeted processes are protected,
# minimum - Modification of targeted policy. Only selected processes are protected.
# mls - Multi Level Security protection.
SELINUXTYPE=targeted
```

รูปที่ 4.3 การเปลี่ยน selinux เป็น Disabled

X

5) ตั้งค่า swappiness ให้น้อยกว่าหรือเท่ากับ 10

ถ้ามากกว่า 10 ต้องทำให้น้อ<mark>ยกว่าหรือเท่ากับ</mark> 10 ในที่นี้ผู้จัดทำการตั้งค่าไว้ที่ 10

#sysctl vm.swappiness=10

<pre>[root@localhost ~]# vi /etc/sysconfig/selinux [root@localhost ~]# vi /etc/sysconfig/selinux [root@localhost ~]# vm_swappiness</pre>	₽ root@localhost:~		_	×
<pre>lost: vm.swappiness: command not found [root@localhost ~]# sysctl vm.swappiness vm.swappiness = 30 [root@localhost ~]# ^C [root@localhost ~]# sysctl vm.swappiness=10 vm.swappiness = 10 [root@localhost ~]# sysctl vm.swappiness vm.swappiness = 10</pre>	<pre>[root@localhost ~]# vi /etc/sysconfig/selinux [root@localhost ~]# vi /etc/sysconfig/selinux [root@localhost ~]# vm.swappiness bash: vm.swappiness: command not found [root@localhost ~]# sysctl vm.swappiness vm.swappiness = 30 [root@localhost ~]# ^C [root@localhost ~]# sysctl vm.swappiness=10 vm.swappiness = 10 [root@localhost ~]# sysctl vm.swappiness vm.swappiness = 10</pre>	EI 7		Ŷ

ร**ูปที่ 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

Proot@localhost:~

root@localhost ~]# vi /etc/rc.local !/bin/bash THIS FILE IS ADDED FOR COMPATIBILITY PURPOSES

It is highly advisable to create own systemd services or udev rules to run scripts during boot instead of using this file.

In contrast to previous versions due to parallel execution during boot this script will NOT be run after all other services.

Please note that you must run 'chmod +x /etc/rc.d/rc.local' to ensure that this script will be executed during boot.

touch /var/lock/subsys/local echo never > /sys/kernel/mm/transparent_hugepage/enabled echo never > /sys/kernel/mm/transparent_hugepage/defrag

รูปที่ 4.5 การตั้งค่า rc.local

X

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/

root@localhost:/opt - C
[root@localhost opt]# tar -xzvf jdk-8u102-linux-x64.tar.gz -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

```
Proot@localhost:/opt
                                                                                 ×
  You could check uidgid reservation validity in
  /usr/share/doc/setup-*/uidgid file
  [ $UID -gt 199 ] && [ "`/usr/bin/id -gn`" = "`/usr/bin/id -un`" ]; then
umask 002
fi
for i in /etc/profile.d/*.sh ; do
    if [ -r "$i" ]; then
        if [ "${-#*i}"
                                  ]; then
               "$i"
        else
               "$i" >/dev/null
        fi
unset i
unset -f pathmunge
export JAVA_HOME=/usr/java/jdk1.8.0_102
PATH=$PATH:$JAVA_HOME/bin
   INSERT
```

ร**ูปที่ 4.7** การใส่ path ให้กับที่อยู่ของJDK

เมื่อเสร็จแล้วให้ทำการreboot เครื่อง

reboot

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

ot@cloudera:/etc/vum.repos.d

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

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

#systemctl start s<mark>ervic</mark>e http

Packages for Cloudera's Distribution for cm, Version 5, on RedHat or CentOS 7 x86_64 mme=Cloudera's Distribution for cm, Version 5 aseurl=http://cloudera.localhost/repo/cm/5 ogkey = http://cloudera.localhost/repo/cm/RPM-GPG-KEY-cloudera

ร**ูปที่ 4.8** ตัวอย่างการแก้ path

4.1.1.5 ตั้งค่า MariaDB

systemctl start mariadb # mysql -uroot -p Enter password: [Enter] ผ่านได้เลย MariaDB [(none)] > use mysql



Copyright (c) 2000, 2017, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]> use mysql Reading table information for completion of table and column names You can turn off this feature to get a quicker startup with -A

Database changed

ariaDB [mysql]> show databases

Database | information_schema | mysql | performance_schema | test | rows in set (0.00 sec)

MariaDB [mysql]>

รูปที่ 4.9 การเข้า mariadb ด้วย command line

้สามารถเปลี่ยนร<mark>หัสขอ</mark>ง Root ด้ว<mark>ย</mark>

MariaDB [mysql]> update user SET PASSWORD=PASSWORD("welcome1")

WHEREUSER='root';

ให้แน่ใจว่าได้บัน<mark>ทึกก</mark>ารเปลี่ยนแ<mark>ป</mark>ลงโดยการอ<mark>อก</mark>กำสั่ง<mark>ต่อไป</mark>นี้

MariaDB [mysql]> flush privileges;

MariaDB [mysql]> update user SET PASSWORD=PASSWORD("welcome1") WHERE USER='root'; Query OK, 4 rows affected (0.00 sec) Rows matched: 4 Changed: 4 Warnings: 0 MariaDB [mysql]> flush privileges; Query OK, 0 rows affected (0.00 sec)

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

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

ุ (u โ ล *ฮั ๅ ะ*

http://localhost:7180 or 127.0.0.1:7180

ผู้จัดทำใช้

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

User: admin

password: admin

Cloudera Mana; × Cloudera Mana; × Cloudera MANAGER

10

<mark>รูปที่ 4.12</mark> หน้าWeb UI cloudera Manager

Log In

เมื่อเข้าไปจะเจอหน้าให้อ่าน license ดังรูปที่ 4.13 ให้เลือก Yes,I accept และกด continue



จะเจอหน้าให้เลือกรูปแบบการใช้งานดังรูปที่ 4.14 และเลือกแบบ Cloudera Express ซึ่ง เป็นแบบ Free

Welcome to Cloudera Mar ×		and a second state of the	and the second		
← → C O Not secure 192.16	8.56.101:7180/cmf/license/wizard?returnUrl=%2Fcmf%	2Fexpress-wizard%2Fwelc	ome#step=selectLicenseStep		or Q \$
cloudera MANAGER					Support •
	Welcome to Cloudera Manager				
	Which edition do you want to deploy?				
	Upgrading to Cloudera Enterprise provides important feature	res that help you manage and mi	onitor your Hadoop clusters in mission-critical e	environments.	
		Cloudera Express	Cloudera Enterprise	Cloudera Enterprise	
	License	Free	60 Days	Annual Subscription	
			After the trial period, the product will continue to	Upload License Key	
			function as Cloudera Express. Your cluster and	Select License File Upload	
				Clouders Enterprise is available in three editions:	
				Basic Edition Flex Edition	
				Cloudera Enterprise	
	Node Limit	Unlimited	Unlimited	Unlimited	
	CDH	~	~	6.1	
	Core Cloudera Manager Features	~	~		
	Advanced Clouders Manager Features		~	1	
	Cloudera Navigator		~		
	Clouders Navigator Key Trustee				
	Clouders Support				
	cioucera support				
	See full list of features available C in Cloudera Express an	d Cloudera Enterprise.			
			00		
	Back			Continue	
	รูปที่ 4.14 1	หน้าเลือกรูข	ปแบบการทำงาน		
	รูปที่ 4.14 า	หน้าเลือกรูว	ปแบบการทำงาน		
	รูปที่ 4.14 า	หน้าเลือกรูง	ปแบบการทำงาน		
	รูปที่ 4.14 า	หน้าเลือกรูข	ปแบบการทำงาน		
	รูปที่ 4.14 1	หน้าเลือกรูข	ปแบบการทำงาน		
	รูปที่ 4.14 1	หน้าเลือกรูข	ปแบบการทำงาน		
	รูปที่ 4.14 1	หน้าเลือกรูข	ปแบบการทำงาน		
	รูปที่ 4.14 1	หน้าเลือกรูข	ปแบบการทำงาน		2 0 0
	รูปที่ 4.14 1	หน้าเลือกรูข	ปแบบการทำงาน		5
	รูปที่ 4.14 1	หน้าเลือกรูร	ปแบบการทำงาน		2 0 0 0
	รูปที่ 4.14 1	หน้าเลือกรูร	ปแบบการทำงาน		2 0 0 0
	รูปที่ 4.14	หน้าเลือกรูร	ปแบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูร	ปแบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูร	ปแบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูร	ປແบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูร	ປແบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูร	ປແบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูข	ปแบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูข	ปแบบการทำงาน		
	รูปที่ 4.14	หน้าเลือกรูข	ปแบบการทำงาน	ECHI	
	รูปที่ 4.14	หน้าเลือกรูข	ปแบบการทำงาน	ECHI	
	รูปที่ 4.14 1 1 1 1 1 1 1 1 1	หน้าเลือกรูข	ປແບບการทำงาน	ECH	
	รูปที่ 4.14 1 1 1 1 1 1 1 1	หน้าเลือกรูข UTE	ປແບບຄາรทຳงาน	ECH	
	รูปที่ 4.14 1	หน้าเลือกรูข UTE	ປແບບຄາະກຳຈານ	ECH	
	รูปที่ 4.14 (NSTIT	หน้าเลือกรูร UTE	ປແບບຄາະກຳຈານ	ECH	

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

continue

	8.56.101:7180/cmf/express-wizard/welcome	९ ☆
cloudera' MANAGER		Support •
	Thank you for choosing Cloudera Manager and CDH.	
	This installer will install Clouder Finness 5 15 0 and enables you to lister choose nackanes for the services helps (there may be some	a license implications)
	Apache Hadoop (Common, HDFS, MapReduce, YARN) Apache Hadoop (Common, HDFS, MapReduce, YARN)	
	Apache Zoakeeper Apache Dozie	
	Apache Hive Hue (Apache licensed) desche Einme	
	 Apache Impala Apache Sentry 	
	Apache Sqoop Cloudera Search (Apache licensed) Apache Spark	
	You are using Cloudera Manager to install and configure your system. You can learn more about Cloudera Manager by clicking on the	Support menu above.
	Before you proceed, be sure to checkout the CDH and Cloudera Manager Requirements and Supported Versione 🗭	
	Supported Operating Systems IP Supported Databases IP	
	Supported JDK Versions	
		Continue
		1 CDU
	JUVI 4.15 Lhank you for choosing Cloudera Mana	dor ond I IIH
	a - · · · · · · · · · · · · · · · · · ·	
1/2		
1,0,0		
1/04		
1/04		
1/04		
1/04		

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



ในที่นี้ผู้จัดทำใส่ Cloudera.localhost และกคSearch จะได้ดังรูปที่ 4.17และให้กดContinue



192,168,56,101

1 hosts scanned, 1 running SSH, Junu Goodh

ร**ูปที่ 4.17** ตั้งค่า Host เมื่อกคsearch

ต่อมาหน้า select repository ให้ตั้งค่าตามนี้

Choose method ให้เถือก Use Parcels

Back

10

Select the specific release of the Cloudera Manager Agent you want to install on your hosts. ให้เถือก 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-KEYcloudera

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



ต่อมาจะเจอหน้า Accept JDK License ดังรูปที่ 4.19เนื่องจากผู้จัดทำได้ให้ทำการติดตั้งแล้ว จึงไม่ต้องติดตั้งให้กดอีกให้กด Continue

Cluster Installation - C		θ
$\leftrightarrow \rightarrow \mathbb{C}$ \bigcirc Not s	ure 192.168.56.101:7180/cml/express-wizard/wizard#step=javaOptionsStep	Q \$
	Cluster Installation	
	Accept JDK License	
	Oracle Binary Code License Agreement for the Java SE Platform Products and JavaFX	*
	ORACLE AMERICA, INC. ("ORACLE"), FOR AND ON BEHALF OF ITSELF AND ITS SUBSIDIARIES AND AFFILIATES UNDER COMMON CON SOFTWARE TO YOU ONLY UPON THE CONDITION THAT YOU ACCEPT ALL OF THE TERMS CONTAINED IN THIS BINARY CODE LICENSI	TROL, IS WILLING TO LICENSE THE E AGREEMENT AND SUPPLEMENTAL
	LICENSE TERMS (COLLECTIVELY "AGREEMENT"). DECASE READ THE AGREEMENT CAREPLAY RELECTING THE "ACCEPT LICENSE BUTTON AND/OR BY USING THE SOFTWARE YOU ACKNOWLEDGE THAT YOU HAVE READ THE TERMS AND AGREE TO THEM. IF YOU, BEHAIE OF A COMPANY OR OTHER LEGAL ENTITY YOU REPRESENT THAT YOU HAVE THE LEGAL BUTTOR TO READ THE FEAL	. AGREEMENT' (OR THE EQUIVALENT) ARE AGREEING TO THESE TERMS ON NITLY TO THESE TERMS (FOULDO NOT
	HAVE SUCH AUTHORITY, OR IF YOU DO NOT WISH TO BE BOUND BY THE TERMS, THEN SELECT THE 'DECLINE LICENSE AGREEMENT MUST NOT USE THE SOFTWARE ON THIS SITE OR ANY OTHER MEDIA ON WHICH THE SOFTWARE IS CONTAINED.	" (OR THE EQUIVALENT) BUTTON AND YOU
	1. DEFINITIONS. "Software" means the software identified above in binary form that you selected for download, install or use (in the ver-	sion You selected for download, install or
	error concertions provided by Oracle, and any user manuals, programming guides and other documentation provided to you by Oracle and singuistic and singuist	value mes, and data mes, any opage of ader this Agreement. "General Purpose . unctions under end user control (such as but
	not specifically limited to email, general purpose internet browsing, and office suite productivity tools). The use of Software in systems of functionality (other than as mentioned above) or designed for use in embedded or function-specific software applications, for example	and solutions that provide dedicated but not limited to: Software embedded in or
	burbled with moust at control systems, whereas module deprotes, whereas namened overse, knows, i visio by built by uso devices, te equipment, printers and storage management systems, and other related systems are excluded from this definition and not licensed un Java technology applets and applications intended to run on the Java Platform, Standard Editon platform on Java-enabled General Pur	der this Agreement. "Programa" means (a) pose Desktop Computers and Servers; and
	(b) JavaFX technology applications intended to run on the JavaFX Runtime on JavaFX-enabled General Purpose Desktop Computers an Install Oracie Java SE Development Kit (JDK 7)	id Servers. "Commercial Features" means
	Check this box to accept the Oracle Binary Code License Agreement and install the JDK. Leave it unchecked to use a currently installed JP	DK.
	Back	Continue
	ราที่ 10 หม้า Accent IDk License	
	and 4.19 That Accept JDK Electise	
Y .		
	VSTITUTE OF	

ต่อมาจะเจอหน้า Singel User Mode ดังรูปที่ 4.20 ให้กดContinue



ต่อมาหน้า Enter Login Credentials ดังผู้จัดทำใช่Root ในการlogin จึงทำการใส่

แก่password SSH Port:22 number of Simultaneous : 10

Cluster Instal	lation - Clou X			θ -
\leftrightarrow \ominus \mathbf{G}	Not secure 192.168.56.101:7180/cmf/exp	oress-wizard/wizard#step=hostCredentialsStep		Q 🕁 👨
	Cluster Ins Enter Login Cr	tallation		
	Root access to password-less	your hosts is required to install the Cloudera packages. This installer will co sudo/pbrun privileges to become root.	onnect to your hosts via SSH and log in either directly as root or as another user with	
	Login To Ali	Hosts As: region of the test of the test of the test of the test of		
	You may conne Authenticatio	ct via password or public-key authentication for the user selected above. n Method: All hosts accept same password All hosts accept same provate key.		
	Enter	Password:		
	Confirm	Password: SSH Port: 22		
	Number of Sin	ultaneous tallations: (Running a large number of installations at once can consume	large amounts of network bandwidth and other system resources)	
	Back		Continue	
		รปที่ 4.21 หน้า Enter I	Login Credentials	
-		u	5	
V				
N 1 4.				
			5	
	· · · / \		SEV 1	

กด Continue เพื่อไปยังหน้า Install agent ดังรูปที่ 4.22



เมื่อติดตั้งเสร็จจะเป็นดังรูปที่ 4.23 สามารถกคดูข้อมูลการติดตั้งได้โดยกดปุ่ม Detail และ กด Continue เพื่อไปยังหน้าต่อไป



จะเจอหน้า Detecting CDH Versions ดังรูปที่ 4.24 ให้กด Continue เพื่อไปยังหน้าต่อไป



ต่อมาจะเป็นหน้า Inspect hosts for correctness ดังรูปที่ 4.25 ให้กด Finish เพื่อไปยังหน้า

ต่อไป

\leftarrow \rightarrow C \odot Not secure	192.168.56.101:7180/cmf/express-wizard/wizard#step=hostInspectorStep	
cloudera MANAGER		Support +
	Olivater lastellaria	
	Inspect hosts for correctness Run Again	
	Validations	
	Insector ran on all 1 hosts	
	 Individual hosts resolved their own hostnames correctly. 	
	No errors were found while looking for conflicting init scripts.	
	No errors were found while checking /etc/hosts. All hosts resolved localhost to 127.0.0.1.	
	All hosts checked resolved each other's hostnames correctly and in a timely manner.	
	Host clocks are approximately in sync (within ten minutes).	
	No users or groups are missing.	
	No conflicts detected between packages and parcels.	
	No kernel versions that are known to be bad are running.	
	No performance concerns with Transparent Huge Pages settings.	
	CDH 5 Hue Python version dependency is satisfied.	
	O hosts are running CDH 4 and 1 hosts are running CDH 5. All checked hosts in each cluster are running the same version of convoluents.	
	All managed hosts have consistent versions of Java.	
	All checked Cloudera Management Daemons versions are consistent with the server.	
	All checked Cloudera Management Agents versions are consistent with the server.	
	Version Summary	
	Back	Finish
л Т	ร ูปที่ 4.25 หน้า Inspect hosts for correctness	
	ร ูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	E J
E TH	รูปที่ 4.25 หน้า Inspect hosts for correctness	E O O
TH N	รูปที่ 4.25 หน้า Inspect hosts for correctness	
THE	รูปที่ 4.25 หน้า Inspect hosts for correctness	с С О О О
THE	รูปที่ 4.25 หน้า Inspect hosts for correctness	
THE	รูปที่ 4.25 หน้า Inspect hosts for correctness	
THE L	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	
	รูปที่ 4.25 หน้า Inspect hosts for correctness	

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

	C Cluster Setup - Cloudera ×			Θ
	← → C ③ Not secure 192.168.5	6.101:7180/cmf/clusters/1/express-add-services/index		a 🛪 ह
	cloudera MANAGER			Support +
		Cluster Setun		
		Select Services		
		Choose a combination of services to install		
		Core Hadoop		
		HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, and Hue		
		HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and HBase		
		Core with Impala HDES YARN (MagReduce 2 Included). ZooKeeper, Oozie, Hive, Hue, and Impala		
		Core with Search	3 3	
		HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Solr		
		HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Spark		
		All Services HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, HBase, Impale	Solr, Spark, and Key-Value Store Indexer	
		Custom Services		
		Choose your own services. Services required by chosen services will automaticall This wizard will also install the Cloudera Management Service. These are a set of	y be included. Flume can be added after your initial cluster has been set up components that enable monitoring, reporting, events, and alerts: these components require	
		databases to store information, which will be configured on the next page.	a contraction of the contraction	
		Back	Continue	
		ราไท้ 1 26 หม้า	Salact Sarvica	
17		3 D /14.20 /141	Select Selvice	
		/Vcz.		

ต่อมาหน้า Assign Roles ดังรูปที่ 4.27 เป็นหน้าให้เถือกว่าจะให้การทำงานแบบไหนติดตั้ง บนเครื่องไหน แต่ในที่นี้ผู้จัดทำมีเครื่องทำงานเพียงเครื่องเดียวค่าที่ตั้งมาเป็นDefualt ของ cloudera จึงไม่ได้ตั้งค่าใดๆเพิ่มเติมและกด Continue เพื่อไปยังหน้าต่อไป

	Cluster Setup - Cloudera ×					Θ
	← → C O Not secure 192.168.5	5.101:7180/cmf/clusters/1/express-add	d-services/index#step=roleAssignmen	sStep		Q #
	Cloudera MANAGER					Support -
		Cluster Setup				
		Assign Roles				
		You can customize the role assignments for performance of your services. Cloudera do	or your new cluster here, but if assignments are es not recommend altering assignments unless	made incorrectly, such as assigning too many you have specific requirements, such as havi	roles to a single host, this can impact the ng pre-selected a specific host for a specific role.	
		You can also view the role assignments by	host. View By Host			
		HDFS				
		Same As DataNode	SecondaryNameNode × 1 New Same As DataNode	Balancer × 1 New Same As DataNode	Select hosts	
		NFS Gateway	DataNode × 1 New			
		Select hosts	cloudera.localhost +			
		Gateway × 1 New	S Hive Metastore Server × 1 New	WebHCat Server	SHiveServer2 x 1 New	
		Same As DataNode	Same As DataNode	Select hosts	Same As DataNode	
		et) Hue				
		Hue Server × 1 New Same As DataNode	H Load Balancer × 1 New Same As DataNode			
		9 Impala				
		Y Impala Catalog Server × 1 New	¥ Impala StateStore × 1 New	¥ Impala Daemon × 1 New		
		Same As DataNode	Same As DataNode	Same As DataNode -		
		Service Monitor × 1 New	Activity Monitor	Host Monitor × 1 New	Event Server × 1 New	
		Back		8060	Continue	
10						
			วูบท 4.2 7 หน่เ	assign roles		
The second se						
		YST				

ต่อมาจะเป็น Setup Database คังรูปที่ 4.28



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;

MariaDB [(none)]> create database hue; Query OK, 1 row affected (0.00 sec) MariaDB [(none)]> create database hive; Query OK, 1 row affected (0.00 sec) MariaDB [(none)]> create database ooserver; Query OK, 1 row affected (0.00 sec) MariaDB [(none)]>

รูปที่ 4.29 สร้าง database

MariaDB [(none)]> flush privileges; MariaDB [(none)]> exit; เพื่อออกจากMariaDB

10

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 เพื่อไปยัง หน้าต่อไป



ต่อมาจะเป็นหน้า Review Changes คังรูปที่ 4.31 การตั้งก่าทั้งหมดจะเป็น Default จึงไม่ ต้องตั้งก่าอะไรเพิ่มเติมและกด Continue

	Cluster Setup - Cloudera ×					Θ -
	← → C O Not secure 192.168.5	5.101:7180/cmf/clusters/1/e	xpress-add-services/index#step=reviewStep			⊶ Q ☆ ₹
	CIOUDEI 3 MANAGER					Support *
		Cluster Setup				
		Review Changes	Cluster 1 > HDES (Service-Wide)		٢	
		dfs.block.size, dfs.blocksize	128 MiB •		U	
		DataNode Failed Volumes	Cluster 1 > DataNode Default Group		0	
		dfs.datanode.failed.volumes.tole rated	. 0			
		DateNode Data Directory	Cluster 1 > DataNode Default Group 🔷		0	
		dfs.datanode.data.dir	/dfs/dn	3 >	Θ	
		NameNode Data Directories	Cluster 1 > NameNode Default Group	CI SI Y	0	
		dfs.namenode.name.dir	/dfs/nn			
		HDFS Checkpoint Directories	Cluster 1 > SecondaryNameNode Default Group		0	
		fs.checkpoint.dir, dfs.namenode.checkpoint.dir	/us/sm			
		Hive Warehouse Directory hive.metastore.warehouse.dir	Cluster 1 > Hive (Service-Wide)		0	
			/user/hive/warehouse			
		Hive Metastore Server Port hive.metastore.port	Cluster 1 > Hive Metastore Server Default Group 9083		0	
		Kudu Service	Cluster 1 > Impala (Service-Wide)		0	
		Back	I.	2 3 4 5 6	Continue	
			_			
			รปที่ 4.31 หน้า R	Review Changes		
1			3 211 1101 111811	terrett entanges		
						(N
					. C	
		VVC				

ต่อมาจะเจอหน้า First Run Command คังรูปที่ 4.32 clouderaจะทำการrun service ที่ผู้จัดทำ เลือกมาทั้งหมด

	Cluster Setup - Clouder: X Cluster Setup - Clouder: X Cluster Setup - Clouder: X Cluster:	- e
	Construction of the second of the secon	
	Cluster Setup	
	First Run Command	
	Status wanning vag au, 400, 400, 400 and 400	
	Show All Steps O Show Only Failed Steps O Show Running Steps	
	Cluster 1 C Aug 30, 401 45 PM Abort	
	Start Cloudera Management Service, ZooKeeper	
	> Ø startHors	
	O Start YABN (MR2 Included)	
	Starthive A Grant measurements	
	>0 startinger etc.	
	Back	
	รูปที่ 1 32 หม้า First Run Command	
A		

Cluster Setup - Cloudera 🗙 Θ → C O Not secure | 192.168.56.101:7180/cmf/clu ← on Q 🕁 👨 Cluster Setup First Run Comm Status O Finished 🚔 Aug 30, 4:01:44 PM O 8.1m Finished First Run of the following servi v Co > 🔿 Aug 30, 4:01:45 P 42.65 . . > Start 01.18s 2.2m 123456 Back รูปที่ 4.33 Run service success T

เมื่อเปิด Service ทั้งหมดจะได้ดังรูปที่ 4.33 และกด Continue ไปยังหน้าต่อไป

ต่อมาเมื่อติดตั้งเสร็จจะเจอหน้<mark>ำ Congratulat</mark>ion ดังรูปที่ 4.34 ให้กด Finish เป็นอันเสร็จ สิ้นการติดตั้ง Cloudera และ CDH


4.1.2 ติดตั้งโปรแกรม StreamSets ถงบน Cloudera Manager

เป็นขั้นตอนการลงโปรแกรม StreamSets ซึ่งเป็นโปรแกรม open source ใช้ทำ ETL ที่มี ความยืดหยุ่นสูง เมื่อต้องการทำETL Big Data นั้นก็ต้องทำให้อยู่ในสภาพแวดล้อมเดียวเพื่อการ ทำงานมีประสิทธิภาพสูงสุดด้วย Cloudera จึงทำการติดตั้ง StreamSets บน Cloudera ซึ่ง Cloudera มีตัวเชื่อมต่ออยู่แล้ว วิธีการติดตั้ง

Document:https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuid e/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

ใปที่ 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

ฐปที่ 4.36 Custom Service Description

นโลยั1ก

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

Settings

Filters

C O Not secure | 192.168.56.101:7

cloudera MANAGER

@ ☆ ₹

ê 43

4.1.2.4 ตรวจสอบ path cloudera parcel

10

C O Not secure | 192.168.56.10

cloudera MANAGER Settings

ตรวจสอบว่า Directory /opt/cloudera/parcel-repo ใด้ใช่จริงหรือไม่ 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

STITUTE O



รูปที่ 4.38 Host_Parcels

จะมี STREAMSETS_DATACOLLECTOR ให้เลือก Distribute คังรูปที่ 4.39



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



ให้เลือก Service StreamSets แล้วกด Continue ดังรูปที่ 4.42



ต่อจะมาหน้าให้เลือก hosts ให้กคที่ Select hosts ดังรูปที่ 4.43

	-		
	← → C ① Not secure 192.16	8.56.101:7180/cmf/clusters/1/add-service/index?serviceType=STREAMSETS#step=roleAssignmentsStep	Q
	cloudera MANAGER		Su
		Add StreamSets Service to Cluster 1	
		Assign roles for our anisotia	
		suffer. You can also view the role assignments by host. View By Host	
		A Data Collector	
		Select hosts Too few hosts assigned, minimum is 1.	
		Back	
		UNA COMME	
		3011 4.43 Idell Assign role for StreamSets	
7-			
- N			
	VI_		
		Werner of V	

เมื่อกดแล้วให้เลือก hosts ที่ต้องการดังรูปที่ 4.44 เสร็จให้กด OK และ กด Continue

Add StreamSets Service to × Θ C O Not secure | 192.168.56.101:7180/cmf/clu: ← Q 🖈 🖑 1 Host Sel CAP CES a DC C SM III NM H RM Сн 00 1-1 of 1 รูปที่ 4.44 Select hosts 10 Cloudera manager จะทำการ run command เพื่อเปิด service เมื่อเสร็จกด continue ดังรูปที่ 4.45 C Add StreamSets Service to × Θ ← → C O Not secure | 192.168.56.101:7180/cmf/clusters/1/add-service/index?serviceType=STREAMSETS#step=comm Q cloudera MANAGER Add StreamSets Service to Cluster 1 First Run Command Status 🗢 Finished 🏥 Aug 28, 5:43:08 PM 📀 24.82s Finished First Run of the following services successfully Completed 2 of 2 step(s). Show Only Failed Steps Aug 28, 5:43:08 PM > 0 ring that the 28. 5:43:08 PM รูปที่ 4.45 Cloudera start service StreamSets

ต่อมาจะเป็นหน้า Congratulation ถือเป็นการเสร็จสิ้นการติดตั้งStreamSets บน Cloudera ดังรูปที่ 4.46

C O Not secure

cloudera MANAGER

T

192.168.56.101:7180/cmf/c

Congratulations!

Add StreamSets Service to Cluster

รูปที่ 4.46 ติดตั้ง StreamSets สำเร็จ

0200

ุกโนโลยั7 พ 0

Q

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 เมื่อเข้า ครั้งแรกจะให้ทำการตั้งsuperuser ทันที



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

⊒ ⊕∪e	Documents -	
	8 8	My document:
 ✔ ■ default Tables 	(1) 🕇 🕂 😂	Ο Ο
I area_callcenter	TE	

ร**ูปที่ 4.4**9 ไอคอนเมนู 3 ขีดของ hue

69



รูปที่ 4.50 Menu_File

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

Θ

07 Q

T

×

E File Browser

C O Not secure | 192.168.56.101

Hue - File Browser

2 4

HUe

รูปที่ 4.51 หน้าเก็บ File hdfs	
NSTITUTE OF	



รูปที่ 4.52 สร้าง File หรือ Directory

4.1.3.3 upload file เข้า HDFS

เมื่อกด upload จะขึ้น popup ให้เราเลือกไฟล์ดังรูปที่ 4.53 สามารถเลือกไฟล์ตามที่ต้องการ



เมื่อกด Select Files จะขึ้นหน้าให้เราเลือกไฟล์ เมื่อเลือกเสร็จกด open เพื่อ upload ดังรูปที่

4.54

C O Not secure	192.168.56.101:8	8888/hue/filebrowser/view=	/user/oracle						o , Q
		X	ou are accessing a non-optimized Hus, please swite	ch to one of the available	addresses: http://clouder	localhost:8889			
	Documents +		Q. Upload to /upor/oraclo						Iabs III
	8		opidad to ruser/oracle			^			
			Select files or drag and drop them	n here					
						_			
		# Home / user /	oracle						
			and the second se						
		Open						×	
			25						
		← → ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑	PC > Desktop			V O Search	Desktop	,	
		Organize • New folder					E • 🗆	1 0	
		Pictures	Name	Date modified	Туре	Size		^ ige 1	of 1 144 44
		This PC	Presentation_Public Dupasakul_A-POST	02/04/2561 16:09	Microsoft PowerPo	3,52,3 ND			
		5 3D Objects	Research	07/03/2561.036	Microsoft PowerPo	4.815 KB			
		Desktop	S 4046851	23/07/2561 22:18	JPG File	692 KB			
		P Documents	service	15/05/2561 10:42	Text Document	1 KB		1	
		Downloads	5 Sourcetree	21/04/2561 20:19	Shortcut	3 KB			
		Nusic.	Sqldeveloper	02/08/2561 15:17	WinRAR archive	410,863 KB			
		Pictures	📕 test	28/06/2561 10:43	JetBrains PyCharm	1 KB			
		The Videos	TNI-Internship-Template	01/10/2561 14:24	Microsoft Word D	8,547 KB	10 March 10		
		IT & OF (C)	2 Wallpaper Engine	14/05/2561 23:51	Internet Shortcut	1 KB			
		User (C)	Whatsup Gold	20/09/2561 17:38	Microsoft PowerPo	4,195 KB			
		LOCAL DISK (D:)	i wifi account starbuck	17/06/2561 15:24	Text Document	1 KB			
		Metwork Y	นายายายายายายายายายายายายายายายายายายาย	05/09/2561 17:18	Microsoft Word D	1,014 KB		×	
		F 11	This Internation Tempolate			All Film			

ร**ูปที่ 4.54** หน้าเลือก File เพื่ออัพโหลด

4.1.3.4 ตั้งค่า Files

สามารถตั้งค่าการ replication,permission,compress และอื่นได้ด้วยการคลิกขวาดังรูปที่

4.55

10



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

หรือสามารถนำข้อมูลเข้าโคยใช้ command Linux

T

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

Proot@cloudera:/opt/data					_
[root@cloudera data]# 1 TNI-Internship-Template [root@cloudera data]# h	.s .docx wadoop fs -copyFromLo	ocal TNI-Internshi	p-Template.docx /u	ser/oracle/TNI	
[root@cloudera data]# [racle			Ο	
Name			Size	User	
■ J				hdfs	
D •		• •		oracle	
			8.3 MB	root	
TNI-Internship-Templa	te.docx		8.3 MB	oracle	
Show 45 Tof 2 items			-ECX		

รูปที่ 4.57 ตัวอย่างการนำข้อมูลเข้าด้วย command line

4.1.4 การทำ ETL ด้วย StreamSets

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

4.1.4.1 ใช้ Browser เข้าไปยัง StreamSets

http://localhost:18630 ในที่นี้ผู้จัดทำใช้ http://192.168.56.101 เมื่อเข้าครั้งแรกจะเจอหน้า ดังรูปที่ 4.58



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

StreamSets Data Collector × +		-
← → C ② Not secure 192.168.56.101:18630		🛍 Q 🖈 🖑 🖪 🌞
StreamSets		💩 # tì # 00
	New Pipeline	×
	Title Pipeline Definition Title	
	Description Pipeline Definition Description	
	Data Collector Pipeline Data Collector Edge Pipeline Microservice Pipeline	
5	Can	cel Save
י ה י	a Try a Haonal	
		$\gamma_{1} > \gamma_{2}$

รูปที่ 4.59 สร้าง pipeline ใหม่ StreamSets

เมื่อใส่ข้อมูลTitle และ Description ก็เลือก Data Collector Pipeline และกค Save จะแสดง หน้าดังรูปที่ 4.60 ส่วนที่ 1 สำหรับว่าง flow การทำงานโดยสามารถนำส่วนที่ 3 ที่เป็น ส่วนของ Stage Library ต่างๆ ลากมาว่างในส่วนที่ 1 ได้ ส่วนที่ 2 คือที่ตั้งก่าของ pipeline และ library ต่าง ๆ

110

Stroom Coto	50.101:18630/collector/pipeline,	ETLBIGUATAD3682 14 1-6620-4946-0471-19082 1803031					1
			C All Channel Court	0 B D C			,
ennes / ETE big Data			W An Changes Saved				
				- 1		Type to search	6
							sc
						Amazon 83 Consum	nei
						Azure IoT/Event CoAP Ser	e E Riv
					-	DEV DEV	,
					(#)	Dev Data Generator Record	do "d
					* -	Dev Data Generator DE	do d
s Sie Data +					+ -	Dev Data Generator DE- Dev Raw Det Source Re Wit	do rd DC im
s Big Data +						Dev Data Generator Dev Ranc DE DE V Dev Raw Data Source F. VV DE	do d
Ng Data + No General Parameters	Notifications	s Custer EMR Test Orign				Dev Data Generator Dev Ranc DEV- Dev Raw Data Source Ra SO Data Dev Snapphot Drector	de de 7 DC http://www.inter- http://wwwww.inter- http://wwwww.inter- http://wwwwww.inter- http://www.inter- http://www.i
tig Data + no General Parameters configuration	Notifications A Error Record Pipeline ID E	B Custer EMR Test Organ		-	······································	Der Cate Der Racht Generater DEF Provinsionen Der Source Provinsionen Der Standen Der Stan	de 7 7 DC rem
tig Data + 10 Gentra Parameters configuration Rates	Notifications A Error Recor Pipeline ID C Title C	B Custer EMR Test Orga TLBg/Datablex2141-e820-494a-b477-196a21a55637 TLBg/Datablex2		7	······································	Der Cate Generalte DEF- Der Recor Der Staben Der Staben	de 7 DC Inth Inth Inth Inth Inth Inth Inth Inth
lig Data + no General Parameters contguestion bases ectory	Patriculions Error Recorr Pipeline ID E Title E Description E	B Cluster EMR Test Origin TITLBIDCM4805482141-4820-454a-b477-156421450507 TITLBID DM4 est ETL BIg D41a				Der Dar Der Kann Generalte DE - V Der Rev Der Store Der	do d
lig Data + no General Parameters configuration ectory	Potitications Contractor Pipeline ID C Title C Description C Labels	B Cluster EMR Test Orga TILBIDGMa00ex2141-e020-494a-b477-196a21a5567 TILBID Data est ETL Big Data			······································	Der Rom Der Ro	do 7 7 CCC REM NY NY
ba Data + no General Parameters configuration Rates 48007	Notifications A Enormacon Pipeline ID E Title E Description E Labels Execution Mode C 2	5 Duster EMR Test Orgin TITLBIJDMaDDva2141-4224-454a-a477-156x21a/35637 TITLBIJDMaD est ETL BIg Data				Der Der Annen Der Annen Der Recent Der Recen	40 d0

รูปที่ 4.60 Pipeline StreamSets

4.1.4.2 การทำ Flow ETL

1) ตั้งค่า pipeline

เมื่อCreate new pipeline มาถึงจะเห็นError เครื่องหมายตกใจสีแดง จะเป็นตัวแจ้งเตือนให้ เราตั้งค่าให้ถูกต้อง ให้เรากดเลือก Error Records ดัง รูปที่ 4.61



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

ุ ุนโล*ย*ั

รูปที่ 4.63 ตั้งค่าที่อยู่เมื่อเขียน File เสร็จ

StreamSets Pipelines / ETL Big Data

TC

2) Origin

TC

nSets Big Data จะเป็นส่วนต้นทางข้อมูลที่เข้ามาในส่วนนี้มี File CSV โคยข้อมูลที่มีคังรูปที่ 4.64

	D'	ouchiti	id.csv - L	.ibreOffic	e Calc			
	<u>F</u> ile	<u>E</u> dit	View	Insert	F <u>o</u> rma	t Styles	Sheet	Data
		• 🖻	•		3 🔞	1 😽 🖳	-	<u>)</u> 🦼
	Libe	eration	Sans	~ 10	~	l a a	a	- 🔳 -
	A1		[~ 5	Σ = Ι	id		
		A		В		С)
	1	id	name					
1	2	101	puchit					
	3	102	wattana	1				
	4	103	max				÷ .	
đ	5	104	viroj	-				
1	6	105	wattana	akornvir	Qİ			
	7	106	puchito					
	-							

รูปที่ 4.64 ข้อมูลทดสอบในFile CSV

ให้ทำการเลือก Directory ในกรอบแคงกคกลิกซ้ายที่ icon เพียงครั้งเคียวจะได้ดังรูปที่ 4.65 เมื่อเลือกแล้วให้ทำการตั้งค่า path ที่อยู่ของ file และประเภทของไฟล์ผู้จัดทำใส่ดังนี้

Files Directory : /u01/bigdata/data File name Pattern : *.csv (เลือกทุกไฟล์ที่มี .csv) และเลือกหัวข้อ Data Format

1

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

Directory 1 -					
Info	General	Files	Post Processing	Data Format	
Configuration			Dat	a Format 📵	Delimited
1 External Libraries			Compressio	n Format 🕕	None
			Delimiter For	mat Type 🙃	Custom
					Use I lander line
			He	ader Line O	will neader Lille
			Allow Extra	Columns 🕕	
			Max Record Lengi	n (cnars) 😈	
			Delimiter (Character 🕕	Tab Semicolon Comma Space Other
			Escape (Character 🕕	Tab Semicolon Comma Space Other \
			Quote 0	Character 🕕	Tab Semicolon Comma Space Other
			Enable c	omments 📵	
			Ignore en	ipty lines 🕕	8
			Root F	ield Type 🕕	List-Map
			Line	s to Skip 🕕	8
			Pars	e NULLs 🕕	
				รป	ที่ 4.66 Data format Delimited
				ଧ	
	•				
3	Proces	ssors			
D.	000000	ore 9	ำหน้าพื่	เปลี่ยว	แปลงปรับประทั่งบล ให้ทำการเลือก store เป็น Proposes 6
ri	ocess	015 1	IINHIN	10 PION	when a hand in the second stage of a Liocessoisan
รูปที่ 4.67	Proce	ssor	S		



เลือก Field Type Converter จะใด้ดังรูปที่ 4.68



เลือก Fields to Convert คังรูปที่ 4.70 และเลือก /id หรือพิมพ์ก็ได้

Convert to Type : INTEGER



รูปที่ 4.71 Hive Metadata

I.

{LOG}

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



		83
	Create a new database	
	No source data	»
	DESTINATION	
	Name Database name A Empty name or invalid characters	
	PROPERTIES	
	Description Description	
	รูปที่ 4.74 Create a new database	
	เมื่อสร้างเสร็จแล้วจะขึ้นคังรูปที่ 4.75	
	I Table Browser	5
	Databases	
	Search for a database X Drop	
	Database Name	তি
y,	default	Õ l
	puchit	
	รูบท 4.75 สร้าง database เรียบรอย	

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

	Directory 1	Field Type Converter	1 re Metadata 1 2	
			Sr.	
			.5-	
Г С	General Hive Table Advanced Data Format			
	JDBC URL	jdbc:hive2://192.168.56.101:10000/puchit		
	JDBC Driver Name	org.apache.hive.jdbc.HiveDriver		
THE REAL	Hadoop Configuration Directory 1 Additional Hadoop Configuration 1	/etc/hive/conf		e Va
	รูป INSTI	ที่ 4.76 ตั้งค่า Hive metadata TUTE OF	ECH	

4) Destinations

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



เลือก Stage Hadoop fs เลือกหัวข้อ General

Stage Library : CDH x.xx.xx

Hadoop FS 1 -							
Info	General	Hadoop FS	Output Files	Late Record	is Data Format		
Configuration				Name	Hadoop FS 1		
External Libraries				Description			
			s	tage Library	CDH 5.13.1		
			Proc	duce Events	•		
			Require	ed Fields 🕕			
							Select Fields Using Pre

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 ของ

ผู้ใช้งาน)

70

Idel timeout : \${1 * SECONDS}

86

		87
3 Output Files Late Record	s Data Format	
File Type 🕚	Text files	•
Files Prefix 🕚	puchit	
Files Suffix 🕚	csv	
Directory in Header 🕚		
Directory Template 🕚	/user/hive/warehouse/puchit.db	
Data Time Zone	+00:00 UTC (UTC)	•
Time Basis 🕕	\${time:now()}	
Max Records in File 🕕	0	
Max File Size (MB) 🕚	0	
ldle Timeout 🕚	\${1 * SECONDS}	

รูปที่ 4.80 Hadoop fs หัวข้อ Output files

หัวข้อ Data format Data Format : Avro Avro Schema Location : In Record Header ต่อมาให้เถือก Stage Hive Metastore หัวข้อ General Stage Library : CDH x.xx.xx หัวข้อ Hive

10

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

				88
	Directory 1	Field Type Converter Hive Metadata 1	Hive Metastore 1	
General	Hive Advanced JDBC URL ① JDBC Driver Name ③ Ittional JDBC Configuration Properties ③	jdbc:hive2://cloudera.localhost:10000/puchit org.apache.hive.jdbc.HiveDriver Property Name Value		
	Hadoop Configuration Directory ① Additional Hadoop Configuration ①	* /etc/hive/conf * รปที่ 4.81 Hive Metastore หัวข้อ	Switch to bu	ilk edit mo
Ŧ				
	CHI INS		TECHT	

4.1.5 ทดสอบความถูกต้องหลังการ โอนข้อมูลเข้า HDFS

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



เมื่อเข้ามาแล้วจะ ได้หน้าดังรูปที่ 4.83 ช่องด้านซ้ายจะเป็น database สามารถกด คลิกเพื่อเข้า ไปดู table ได้ส่วนช่องด้านขวาเป็นที่ใส่ Code SQL เพื่อ Query ข้อมูล



4.1.6 การตั้งเวลาในการทำงานให้ StreamSets

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



เมื่อกคเข้ามาจะเห็นDocument กคเลือกแล้วเปลี่ยนเป็น Actionคังรูปที่ 4.86

My Workflow Add a description.

Drop your action here

DOCUMENTS -

TC

J.

E

รูปที่ 4.86 Ozzie เปลี่ยน Document เป็น actions

เมื่อเลือกแล้วจะเห็นเป็นดังรูปที่ 4.87 ให้เลือก shell ลากไปที่ drop your action





รูปที่ 4.89 Script การสั่งstart StreamSets

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


เมื่อได้ดังรูปที่ 4.91 แล้วให้กด รูปsave ในกรอบสีแดง



ให้ตั้งชื่อ schedule และเลือก Choose a workflow คังรูปที่ 4.93

Oozie Editor

start streamsets

Add a description...

Which workflow to schedule? a workflow...

Choose a workflow...

TC

รูปที่ 4.93 เลือก schedule

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

	O Oozie Editor
	start streamsets
	Add a description
	Which workflow to schedule?
	Start StreamSets
	How often?
	Every day at 13 : 25
	₽Hide
	Advanced syntax
	Timezone Asia/Bangkok •
	From 🛗 2018-10-04 🖸 23:13
	To 🗰 2018-10-11 💿 23:13
	Darameters
	+ Add parameter
7-	
	รูปที่ 4.94 ตั้งค่า schedule
	u u
T	
	NCS. OF Y

บทที่ 5 บทสรุปและข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน

จากการที่ได้ศึกษาและทคลองทำ Big Data เพื่อการดึงและรวบรวมข้อมูลขนาคใหญ่เพื่อ นำมาประมวลผลให้เกิดประโยชน์และมีประสิทธิภาพสูงสุดด้วยเกรื่อง Bigdata ซึ่งได้ผลลัพธ์ดังนี้

5.1.1 สามารถติดตั้งระบบ Big Data และเครื่องมือต่างๆที่ใช้ในระบบได้สำเร็จ

5.1.2 สามารถนำข้อมูลเข้าฐานข้อมูลของBig Data ได้สำเร็จ

5.1.3 สามารถทำการดึงข้อมูลจากแหล่งข้อมูลมาปรับปรุงและนำเข้าฐานข้อมูล Big Data ได้สำเร็จ

5.1.4 สามารถกำหนดเวลาการทำงานโดยใช้เครื่องมือ Oozie ได้สำเร็จ

5.1.5 สามารถทำการทคสอบและได้ผลลัพธ์ตามที่กาคหวังไว้

จากผลลัพธ์ทั้งหมด สามารถสรุปได้ว่า การติดตั้งระบบ Cloudera Manager ซึ่งเป็นระบบ จัดการเครื่องมือ Big Data สามารถตอบสนองต่อความต้องการของบริษัท และเครื่องมือ StreamSets สามารถทำการดึงข้อมูลจากแหล่งข้อมูล ปรับปรุงข้อมูล นำเข้าฐานข้อมูลของ Big Data ได้อย่างครบถ้วนตามที่ต้องการ พร้อมทั้งสามารถนำเทคโนโลยีทั้งหมดนี้ไปต่อยอดและพัฒนาใน อนาคตได้

5.2 แนวทางกา<mark>ร</mark>แก้ไ<mark>ขปัญ</mark>หา

G

ปัญหาที่พบในระหว่างการศึกษาและติดตั้งระบบ Big Data นั้นส่วนแรกจะเป็นส่วนของ การทำความเข้าใจในทฤษฎีและภาพรวมของโครงสร้าง Big Data เนื่องจากระบบของ Big Data มี โครงสร้างหลายขั้นตอนพอสมควร ซึ่งทำให้การทำความเข้าใจกระบวนการศึกษาเพื่อใช้งาน เบื้องต้น มีความเข้าใจได้ยาก แนวทางการแก้ปัญหานี้กือ เรียนรู้ผ่านเว็บไซต์ที่มีผู้พัฒนาได้มาเขียน ไว้ หรือศึกษาจากเอกสารหลักของเรื่องนั้นๆ ส่วนอีกปัญหาที่พบคือ การติดตั้งโปรแกรม Cloudera Manager นั้นหากตั้งก่าของโปรแกรมต่างๆ ไม่ถูกต้อง หรือไม่เหมาะสมแก่การติดตั้งในส่วนต่อไป ก่อนข้างมีความยุ่งยาก และอาจเกิดปัญหากรเข้าใช้งานไม่ได้ ซึ่งแนวทางการแก้ปัญหากือ ศึกษาว่า Cloudera Manager จะต้องการสิ่งใดบ้างเพื่อที่จะได้วางแผนการติดตั้งระบบได้ และในการตั้งก่า ต่างๆกวรรอบคอบละเอียดที่สุด

5.3 ข้อเสนอแนะจากการดำเนินงาน

10

5.3.1 ควรมีพื้นฐานในการใช้ระบบปฏิบัติการ Linux ทางด้านการใช้คำสั่งและการตั้งค่า ระบบพื้นฐานของเครื่อง

5.3.2 ควรมีพื้นฐานและความเข้าใจในการติดตั้ง Package เนื่องจาก Cloudera Manager ต้องการ Package ก่อนการติดตั้งหลาย Package ด้วยกัน

5.3.3 ควรมีพื้นฐานการใช้งานและติดตั้ง MySQL เนื่องจากเครื่อง Big Data บ้างเครื่องมือ จำเป็นต้องมีที่เก็บข้อมูล

5.3.4 ในการทำงานเป็นต้องหาข้อมูลเพื่อเตรียมกวามพร้อมก่อนลงมือและต้องวางแผน ก่อนการทำงานเสมอ

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 ดุลาคม 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]

MSTITUTE OF T



การเตรียม Virtual machine โดยใช้ Oracle VM VirtualBox

การเตรียม Virtual Machine โดยใช้ Oracle VMware VirtualBox



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

		?	×
← Creat	e Virtual Machine		
Name	and operating system		
Please c and sele it. The n identify t	hoose a descriptive name for the ne ct the type of operating system you name you choose will be used through this machine.	ew virtual i intend to i ghout Virtu	machine install on ialBox to
Name:	Big Data		
Type:	Linux		64
Version:	Other Linux (64-bit)	Ũ	7
	Expert Mode Next		Cancel

รูปที่ ก.2 การติดตั้ง VMware (2)

- เถือดขนาด RAM การติดตั้ง Cloudera อย่างน้อยต้องมี 32 GB แต่เนื่องจากเอาให้ใช้ได้ เบื้องต้นจึงตั้ง 10240 mb

T

? ×
← Create Virtual Machine
Memory size
Wentory size
Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.
The recommended memory size is 512 MB.
MB 🖨 معظمه 🖬
4 MB 16384 MB
Next Cancel

รูปที่ ก.3 การติดตั้ง VMware (3)

- เถือก Create a virtual gard disk now

- เลือก VDI

					2	\sim
					1	~
100						
← Cr	eate Virtua	Machine				
Har	d disk					
1 Ion	a ansie					
If yo	u wish you c	an add a vir	tual hard d	lisk to the	new m	nachine.
You	can either cr	eate a new	hard disk f	ile or sel	ect one	from
the l	ist or from a	nother locat	ion using t	he folder	icon.	
If yo	u need a mo	re complex	storage se	t-up you	can skip	this
step	and make th	e changes f d	to the mac	hine setti	ngs onc	e the
mac	nine is create	.u.	-			
The	recommende	ed size of th	e hard disk	c is 8.00	GB.	
0	o not add a v	/irtual hard	disk			
0 0	Create a virtu	al hard disk	now			
0	Jse an existin	q virtual ha	rd disk file			
	O Claudara	ContOSZ	di (Normal	50.00.0	D)	
	25 Cloudera	_CentOS7.V	di (Normai	, 50.00 G	в)	×
			Cre	ate	Ca	ncel
				_		
	19	9	e			
Ţ						
						? ×
Create	Virtual Hard	Disk				? ×
← Create	: Virtual Hard	Disk				? ×
← Create	e Virtual Hard	Disk	-			? ×
← Create	e Virtual Hard	Disk				? ×
← Create Hard d	Virtual Hard isk file type	Disk e f file that you	would like to	o use for th	e new v	? ×
← Create Hard d Please th disk, if yo this settin	Virtual Hard isk file typ oose the type o	Disk e f file that you	would like to	o use for th	ie new v ware you	? ×
← Create Hard d Please ch disk. If yo this settin () ♥ voi (\	Virtual Hard isk file type oose the type of u do not need i g unchanged. VirtualBox Disk I	Disk e f file that you o use it with mage)	would like tr	o use for th	te new v ware you	? ×
← Create Hard d Please ch disk. ff yo this settin ● VoI (\ ○ VHO ()	Virtual Hard isk file type u do not need g unchanged. irtualBox Disk I Virtual Hard Disk	Disk e f file that you o use it with mage) k)	would like tr	o use for th	ie new v ware you	? X
← Create Hard d Please ch disk. If yo this settin ● VDI (v ○ VHD (v ○ VHD (v	Virtual Hard isk file type oose the type of u do not need ig unchanged. Virtual Box Disk I Virtual Hard Disk Virtual Hard Disk	Disk e f file that you o use it with mage) k) e bisk)	would like tr	a use for th	ie new v ware you	? X
← Create Hard d Please ch disk. If yo this settin ● VDI (v ○ VHD (i ○ VMDK	Virtual Hard isk file type oose the type of g unchanged. VirtualBox Disk I Virtual Hard Dis (Virtual Hard Dis	Disk e f file that you o use it with mage) k) e Disk)	would like tr	o use for th	ie new v	? >
← Create Hard d Please ch disk, if yo this settin ● VDI (V ○ VHD (' ○ VMDK	Virtual Hard isk file type oose the type of u do not need ig unchanged. VirtualBox Disk I Virtual Box Disk I Virtual Machir	Disk e f file that you io use it with mage) k) e Disk)	would like tr	o use for th	ie new v	? ×
← Create Hard di Please ch disk. If yo this settin ● VDI (V ○ VHD () ○ VMDK	• Virtual Hard isk file type oose the type of u do not need n g unchanged. VirtualBox Disk I Virtual Hard Dis (Virtual Machir	Disk e f file that you o use it with mage) k) e Disk)	would like tr	o use for th	ie new v ware you	? ×
← Create Hard d Please ch disk. If yo this settin ● VDI (v ○ VHD () ○ VHD ()	• Virtual Hard isk file type w do not need g unchanged. VirtualBox Disk I Virtual Hard Dis (Virtual Machir	Disk e f file that you o use it with mage) k) e Disk)	would like to	o use for th	ie new v ware you	? ×
 Create Hard di Please ch disk. fly and vol (v vol (v) vol (v)<td>: Virtual Hard isk file type u do not need ig unchanged. VirtualBox Disk 1 Virtual Hard Dis (Virtual Machir</td><td>Disk e f file that you o use it with mage) k) e Disk)</td><td>would like to</td><td>a use for th</td><td>ie new v ware you</td><td>? X</td>	: Virtual Hard isk file type u do not need ig unchanged. VirtualBox Disk 1 Virtual Hard Dis (Virtual Machir	Disk e f file that you o use it with mage) k) e Disk)	would like to	a use for th	ie new v ware you	? X
 Create Hard d Please ch disk. <i>B</i>'y vb1 (v VHD (v VHD (v VHD (v VMDK 	Virtual Hard isk file typ u do not need g unchanged. VirtualBox Disk I Virtual Hard Dis (Virtual Machir	Disk e f file that you o use it with mage) k) e Disk)	would like to	a use for th	ie new v ware you	? ×
← Create Hard d Please ch disk. If yo this settin ● VDI (\ ○ VHD (\ ○ VHD K	Virtual Hard isk file type oose the type o u do not need i gunchanged. VirtualBox Disk I Virtual Hard Dis (Virtual Machir	Disk e f file that you o use it with mage) k) e Disk)	would like to	o use for th	ie new v ware you	? X
← Create Hard d Please ch disk. fry otol (\ ○ VHD (\ ○ VHD (\ ○ VHD K	Virtual Hard isk file type oose the type of u do not need i g unchanged. VirtualBax Disk I Virtual Hard Dis (Virtual Machir	Disk e f file that you o use it with mage) k) e Disk)	would like to	o use for th	ie new v ware you	? X
← Create Hard d Please ch disk. If yo this settin ● VDI (\ ○ VHD (' ○ VMDK	Virtual Hard isk file type oose the type of u do not need ig unchanged. VirtualBax Disk I Virtual Hard Dis (Virtual Machir	Disk e f file that you o use it with mage) k) e Disk)	would like to other virtuals	o use for th zation soft	ie new v wware you	? ×

ร**ูปที่ ก.5** การติดตั้ง VMware (5)

- เลือก Fixed size

← Create Virtual Hard Disk

Storage on physical hard disk

Please choose whether the new virtual hard disk file should grow as it is used (dynamically allocated) or if it should be created at its maximum size (fixed size).

A **dynamically allocated** hard disk file will only use space on your physical hard disk as it fills up (up to a maximum **fixed size**), although it will not shrink again automatically when space on it is freed.

A **fixed size** hard disk file may take longer to create on some systems but is often faster to use.

a

Next Cancel

Dynamically allocated
 Fixed size

ร**ูปที่ ก.6** การติดตั้ง VMware (6)

- เลือก size hard disk 40 GB

TC

	?	\times
← Create Virtual Hard Disk		
File location and size		
Please type the name of the new virtual hard disk file i on the folder icon to select a different folder to create	nto the box below the file in.	or click
Big Data		
Select the size of the virtual hard disk in megabytes. T	his size is the limi	t on the
amount of file data that a virtual machine will be able to	to store on the ha	rd disk.
		40.00 GB
4.00 MB	2.00 TB	
	Create	Cancel

รูปที่ ก.7 การติดตั้ง VMware (7) STITUTE OF

- รอการVMware สร้างที่เก็บข้อมูล



108

มากผนวก ข. การติดตั้ง CentOS 7 Linux

CAN INSTITUTE OF TECH

การติดตั้ง CentOS 7 Linux

- เถือก CentOS 7 iso ไฟล์ที่เตรียมไว้และกค Start



รูปที่ ข.1 การติดตั้ง Linux (1)

- เลือก install CentOS 7



รูปที่ ข.2 การติดตั้ง Linux (2)

- เลือกภาษาอังกฤษ (USA)

10



Input Devices Help INSTALLATION SUMMARY CENTOS LINUX 7 INSTALLATION 🖾 us Help! CentOS LOCALIZATION DATE & TIME Americas/New York timezone KEYBOARD English (US) LANGUAGE SUPPORT English (United States) SOFTWARE INSTALLATION SOURCE SOFTWARE SELECTION \bigcirc SYSTEM Checking software depende KDUMP Kdump is enabled INSTALLATION DESTINATION S SECURITY POLICY NETWORK & HOST NAME 🔰 🕤 🌬 🥭 🧰 🜌 🚰 🛄 🍼 🚳 Right Cir

รูปที่ ข.4 การติดตั้ง Linux (4)

- เลือก Asia / Bangkok และกด Done

10



รูปที่ ข.5 การติดตั้ง Linux (5)

- กลับมาหน้าหลักแล้วเลือก Software Selection ให้เลือก Server with GUI และกด



รูปที่ ข.6 การติดตั้ง Linux (6)

- กลับมาหน้าหลักเลือก Installation Destination เพื่อจัดการเรื่องdisk



ร**ูปที่ ข.7** การติดตั้ง Linux (7)

- กลับมาหน้าหลักจะสามาถกด Begin Installation ได้และจะเริ่ม install

TC



รูปที่ ข.8 การติดตั้ง Linux (8)

- ให้เลือก root และตั้งpassword

TC



รูปที่ ข.9 การติดตั้ง Linux (9)

.....

รูปที่ ข.10 การติดตั้ง Linux (10) ////STITUTE OF

ประวัติผู้จัดทำโครงงาน

ชื่อ – สกุล

นายภูชิสส์ วัฒนกรวิโรจน์

วัน เดือน ปีเกิด

26 ธันวาคม 2539

ประวัติการศึกษา

ระดับประถมศึกษา

โรงอุคมศึกษาลาคพร้าว

ระดับมัธยมศึกษาตอนต้น โรงเรียนอุดมศึกษาลาดพร้าว

ระดับมัธยมศึกษาตอนปลาย โรงเรียนมัธยมวัดบึงทองหลาง

ระดับอุดมศึกษา

คณะเทคโนโลยีสารสนเทศ สาขาเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีไทย – ญี่ปุ่น

ทุนการศึกษา

16

- ไม่มี -

ประวัติการฝึกอบรม Training Pre-Cooperative Education โกรงการสหกิจฯบริษัทเอ-โฮสต์ จำกัด

<mark>ผลงานที่ได้รับการตีพิมพ์ - ไ</mark>ม่มี -

STITUTE O

