



ขอบข่ายงานการวิเคราะห์ข้อมูลโดยใช้แพลตฟอร์มข้อมูลขนาดใหญ่

คลอเดราและแรพพิทไมน์เนอร์ราดูป

A Data analysis framework using Cloudera big data

platform and Rapidminer Radoop

นายเนติพงษ์ เทพวุฒิสถาพร

Netiphong Thepwuttisathaphon

นายสิทธิพร คุณกรธรรมนพ

Sittiporn Kunnakorntammanop

โครงการนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

มหาวิทยาลัยศรีนครินทรวิโรฒ ปีการศึกษา พ.ศ. 2560



คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

โครงการ	A Data analysis framework using cloudera big data platform and rapid miner radoop ขอบข่างานการวิเคราะห์ข้อมูลโดยใช้แพลตฟอร์มข้อมูลขนาดใหญ่ คลอเดราและแรพพิทไมน์เนอร์ราดอป		
นิสิต	นายเนติพงษ์	เทพวุฒิสถาพร	58102010810
	นายสิทธิพร	คุณกรธรรมนพ	58102010829
ปริญญา	วิทยาศาสตรบัณฑิต (วท.บ.)		
ภาควิชา	วิทยาการคอมพิวเตอร์		
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.ศุภชัย ไทยเจริญ		

ลงชื่อ

(ผู้ช่วยศาสตราจารย์ ดร.ศุภชัย ไทยเจริญ)

บทคัดย่อ

ข้อมูลขนาดใหญ่ไม่ได้มีเพียงความต่างกันของข้อมูลและการเกิดขึ้นอย่างรวดเร็วของข้อมูลเท่านั้น แต่ยังมีข้อมูลในปริมาณมากอีกด้วย ข้อมูลประเภทนี้ไม่สามารถจัดการได้โดยวิธีการและเครื่องมือทั่วไปเช่นฐานข้อมูลเชิงสัมพันธ์ ถึงแม้ว่าจะมีเครื่องมือของข้อมูลขนาดใหญ่อย่างเช่น **Hadoop, Hive, Spark** ถูกพัฒนาขึ้นมาและประสบความสำเร็จในการนำไปใช้จัดการกับข้อมูลขนาดใหญ่ แต่ยังคงยากที่จะเรียนรู้แม้จะเป็นผู้เชี่ยวชาญก็ตาม ดังนั้น การนำเสนอขอบข่ายการทำงานในการจัดการกับข้อมูลขนาดใหญ่และการวิเคราะห์ข้อมูลที่ง่ายและเป็นมิตรต่อผู้ใช้งานจึงได้เกิดขึ้น ขอบข่ายการทำงานนี้เป็นการรวมกันของแพลตฟอร์มข้อมูลขนาดใหญ่ **Cloudera** กับ **RapidMiner Radoop** ส่วนของ **Cloudera** คือการรวมกันของซอฟต์แวร์ที่สำคัญในการจัดการกับข้อมูลขนาดใหญ่เช่น **Hadoop, Hive, Spark** และ **Zookeeper** เข้าด้วยกันและมี **Cloudera Manager** เป็น GUI (Graphical User Interface) เพื่อใช้ในการจัดการซอฟต์แวร์เหล่านั้น มากไปกว่านั้น **RapidMiner Radoop** ยังทำให้เห็นขั้นตอนการทำงานอย่างเป็นรูปเป็นร่างและยังมีชุดคำสั่งที่ช่วยให้การทำวิทยาศาสตร์ข้อมูล, การวิเคราะห์ข้อมูลใหญ่นั้นง่ายขึ้นจึงทำให้ผู้ไม่มีประสบการณ์หรือผู้ที่มีประสบการณ์เพียงเล็กน้อยสามารถเรียนรู้และจัดการด้วยตนเองได้

Abstract

Big data is not only heterogeneous and fast generated but also large in volume. This type of data cannot be handled by traditional techniques and tools such as relational databases. Although some existing big data software tools such as Hadoop, Hive, and Spark have been developed and successfully used for managing big data, their complexity requires a very steep learning curve, even for technical professionals. Accordingly, to encourage data analytic community, a simplistic, user-friendly framework for analyzing big data is presented. The framework combines Cloudera big data platform with RapidMiner Radoop data science software suite. Cloudera platform unifies a number of key big data software components, such as Hadoop, Hive, Spark, and Zookeeper into a single unit and provides GUI Cloudera Manager for managing them. Moreover, RapidMiner Radoop provides a visual workflow and a set of operators that facilitates data science and big data analysis tasks much easier with non- or minimal-programming effort. A setup guideline is presented in this paper and examples of data analysis using machine learning are demonstrated and implemented.

ประกาศคุณูปการ

ขอบข่ายงานการวิเคราะห์ข้อมูลโดยใช้แพลตฟอร์มข้อมูลขนาดใหญ่คลอเดราและแรพิดไมน์เนอร์ราดูป (A Data Analysis Framework using Cloudera Big Data Platform and RapidMiner Radoop) ผู้วิจัยได้มุ่งมั่นศึกษาและค้นคว้าอย่างต่อเนื่องจนได้รับความรู้และความเข้าใจจนสามารถพัฒนาออกมาเป็นระบบต้นแบบ Big Data Analytics Platform ได้ โครงการนี้สามารถสำเร็จลุล่วงเพราะได้รับความเมตตาและความช่วยเหลือจากบุคคลหลาย ๆ ท่านตั้งแต่เริ่มต้นจนเสร็จสิ้นโครงการ คณะผู้วิจัยขอขอบพระคุณบุคคลทุกท่านที่ให้ความรู้ คำแนะนำ แนวทางการแก้ไขปัญหา และให้ความอนุเคราะห์เพื่อการศึกษาโครงการนี้

ขอขอบพระคุณ ผศ.ดร.ศุภชัย ไทยเจริญอาจารย์ในที่ปรึกษาโครงการผู้ซึ่งให้คำแนะนำ คำปรึกษา ความใส่ใจ และความช่วยเหลืออย่างเต็มที่รวมไปถึงการตรวจสอบ แก้ไข ตลอดจนให้คำปรึกษาในทุก ๆ รายละเอียดของขั้นตอนการทำงานจนโครงการบรรลุวัตถุประสงค์ตามเป้าหมายที่ตั้งไว้

นายเนติพงษ์ เทพวุฒิสถาพร

นายสิทธิพร คุณกรธรรมนพ

สารบัญ

บทคัดย่อ.....	ค
Abstract.....	ง
ประกาศนุญการ	จ
สารบัญ.....	ฉ
สารบัญรูป.....	ญ
บทที่ 1	1
บทนำ	1
1.1 ความสำคัญและที่มาของปัญหา	1
1.2 วัตถุประสงค์ของโครงการ	2
1.3 ขอบเขตของโครงการ	2
1.4 ประโยชน์ที่คาดว่าจะได้รับ	2
บทที่ 2	3
องค์ความรู้ที่เกี่ยวข้อง.....	3
2.1 งานวิจัยที่เกี่ยวข้อง	3
Tweet Analysis: Twitter Data processing Using Apache Hadoop.....	3
Data Analysis using Hadoop MapReduce Environment.....	3
Sentiment Analysis of Facebook Data using Hadoop based Open Source Technologies.....	3

2.2 ทฤษฎีและแนวคิดที่ใช้ในการทำโครงการงาน	4
2.1.1 Big Data.....	4
2.2.2 Big data platform	4
2.2.3 Cloudera.....	5
2.2.4 Hadoop.....	6
2.2.5 Node.....	7
2.2.6 HDFS (Hadoop Distributed File System).....	8
2.2.7 Yarn (Yet Another Resource Negotiator)	8
2.2.8 MapReduce	10
2.2.9 Hive.....	10
2.2.10 Spark.....	12
2.2.11 ZooKeeper	13
2.2.12 RapidMiner	13
2.2.13 Radoop.....	14
บทที่ 3	15
วิธีการดำเนินโครงการงาน	15
3.1 วิธีดำเนินการวิจัยและวิธีการประเมินผล.....	15
วิธีการดำเนินการวิจัย	15
วิธีการประเมินผล	15
3.2 แผนการดำเนินงานตลอดโครงการงาน.....	16

3.3 อุปกรณ์และเครื่องมือที่ใช้ในโครงงาน	17
3.3.1 ฮาร์ดแวร์ (Hardware).....	17
3.3.2 ซอฟต์แวร์ (Software)	17
3.4 ขั้นตอนการดำเนินงาน.....	18
3.4.1 ศึกษาการทำงานเบื้องต้นของ Cloudera Quickstart บน Virtual Machine	18
3.4.2 ทำการติดตั้งระบบปฏิบัติการ CentOS.....	19
3.4.3 ติดตั้ง Cloudera Manager.....	19
3.4.4 ทำการเพิ่ม Namenode และ Datanode.....	20
3.4.5 ทำการติดตั้ง Cloudera Agent ใน namenode และ Datanode	20
3.4.6 ทำการ Download และ Distribute Parcels ให้แก่ Node ต่างๆ.....	20
3.4.7 ทำการเลือก service ต่างๆ ที่ต้องการใช้งาน.....	21
3.4.8 เปิดใช้งาน node ต่างๆ.....	22
3.4.9 ติดตั้ง Service ให้ node ต่างๆ.....	23
3.4.10 เปิดใช้งาน Cloudera Manager และ Service ต่างๆ	23
3.4.11 เช็สถานะการทำงานของ node ต่างๆ.....	24
3.4.12 ติดตั้ง RapidMiner	25
3.4.13 ติดตั้ง RapidMiner Radoop และ ทำการเชื่อมต่อระหว่าง rapidMiner Radoop กับ Cluster.....	25
3.4.14 ทดสอบการอ่านไฟล์	26
3.4.15 ผลลัพธ์จากการทดสอบการอ่านไฟล์	26
3.4.16 ทำการทดสอบระบบโดยการทำ Decision Tree.....	27

3.4.17 ผลลัพธ์จากการทำ Decision Tree.....	27
3.4.18 ตรวจสอบการทำงานของ Cluster	28
3.4.19 ทำการ Classification Model	28
3.4.20 ทำ Classification Model ในส่วน Cross Validation และ Process Document form Data	29
3.4.21 ข้อมูลที่ได้จากการทำ Classification Model	29
3.4.22 ประสิทธิภาพที่ได้จากการทำ Classification Model	30
3.4.23 ทำ Classify Tweet	30
3.4.24 ข้อมูลของ Tweet ที่ได้ดิงมา	31
3.4.25 ผลลัพธ์ของการทำ Classify Tweet	31
บทที่ 4	32
ผลการดำเนินโครงการ.....	32
4.1 ผลการดำเนินงาน	32
บทที่ 5	33
สรุปผล อภิปรายผล และข้อเสนอแนะ	33
5.1 สรุปผลการดำเนินโครงการ	33
5.2 ปัญหาและอุปสรรคในการดำเนินโครงการ	33
5.3 ข้อเสนอแนะ	33
จากปัญหาข้างต้น การติดตั้งระบบบนเครื่อง Namenode ควรมีหน่วยความจำเข้าถึงโดยสุ่ม (RAM) ที่มากกว่าเดิม เพื่อให้การใช้งานและการประมวลผลเป็นไปอย่างราบรื่น.....	33
แหล่งอ้างอิง	34

สารบัญรูป

รูปที่ 2-1 Big Data Platform	4
รูปที่ 2-2 Hadoop Architecture.....	6
รูปที่ 2-3 HDFS Architecture.....	8
รูปที่ 2-4 YARN Architecture	9
รูปที่ 2-5 Map Reduce	10
รูปที่ 2-6 Making Apache Hive 100x faster.....	11
รูปที่ 2-7 Spark Architecture	12
รูปที่ 2-8 RapidMiner.....	13
รูปที่ 2-9 Radoop	14

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหา

เนื่องจากปัจจุบันนี้การทำงาน ติดต่อ สื่อสาร หรือธุรกรรมออนไลน์เป็นที่นิยมและทุกคนสามารถเข้าถึงได้ ทำให้การเติบโตของข้อมูลนั้นเป็นไปอย่างไม่หยุดยั้ง ทำให้มีข้อมูลปริมาณมหาศาลเกิดขึ้นในแต่ละวัน ซึ่งเป็นเรื่องที่สามารถเกิดขึ้นได้ในทุกองค์กร บางองค์กรมีข้อมูลปริมาณมากในครอบครองแต่ไม่สามารถบริหารจัดการและวิเคราะห์ข้อมูลเหล่านั้นอย่างมีประสิทธิภาพได้

การจัดการกับข้อมูลเหล่านั้นให้เกิดประโยชน์สูงสุดจึงเป็นสิ่งสำคัญ แต่สำหรับการจัดการกับข้อมูลปริมาณมหาศาลนั้นจำเป็นต้องมีทรัพยากรคอมพิวเตอร์ในระดับสูง ในบางองค์กรอาจไม่มีอุปกรณ์เหล่านั้นในครอบครอง

ดังนั้นกลุ่มของผู้วิจัยจึงมีความสนใจที่จะศึกษาและพัฒนาระบบที่สามารถจัดการกับข้อมูลปริมาณมากได้โดยใช้ทรัพยากรคอมพิวเตอร์ที่มีอยู่ทั่วไป นำสร้างระบบที่ร่วมกันประมวลผลได้

1.2 วัตถุประสงค์ของโครงการ

- เพื่อศึกษาหลักการทำงานของ Big Data Analytics Platform สำหรับใช้วิเคราะห์และจัดการข้อมูลขนาดใหญ่
- เพื่อศึกษาการใช้งาน RapidMiner สำหรับการวิเคราะห์ข้อมูลร่วมกับ Big Data Platform
- เพื่อสร้างระบบต้นแบบของขอบข่ายงานการวิเคราะห์ข้อมูลขนาดใหญ่

1.3 ขอบเขตของโครงการ

- ได้ระบบต้นแบบสำหรับวิเคราะห์ข้อมูลขนาดใหญ่
- ระบบประกอบไปด้วยคอมพิวเตอร์ 3 เครื่อง 1 NameNode และ 2 DataNode
- ใช้ Cloudera สำหรับสร้าง Big data analytics platform
- ใช้ Rapidminer Studio สำหรับการทำ Data Analysis

1.4 ประโยชน์ที่คาดว่าจะได้รับ

- ได้ระบบต้นแบบที่สามารถวิเคราะห์ข้อมูลขนาดใหญ่
- สามารถเรียนรู้วิธีการสร้างและวางระบบเพื่อนำไปพัฒนาและต่อยอด

บทที่ 2

องค์ความรู้ที่เกี่ยวข้อง

2.1 งานวิจัยที่เกี่ยวข้อง

ชื่อเรื่อง : Tweet Analysis: Twitter Data processing Using Apache Hadoop

ชื่อผู้แต่ง : Manoj Kumar Danthala

บทความนี้นำเสนอเกี่ยวกับการนำข้อมูล Tweet จาก Twitter มาเพื่อทำการวิเคราะห์เพื่อนำมาใช้ประโยชน์ เช่น การนำมาหาว่าผู้ใช้งานคนใดมีการ Tweet หรือ Retweet เป็นจำนวนเท่าใด, บัญชีใดมีการติดตามมากที่สุด, นำมาใช้เพื่อวิเคราะห์ติดตามธุรกิจและองค์กรต่างๆ และทำให้นักวิจัยสามารถนำข้อมูลที่ผ่านมาจากการจัดเก็บและทำการ MapReduce ไปใช้ได้อย่างมีประสิทธิภาพเป็นต้น

ชื่อเรื่อง : Data Analysis using Hadoop MapReduce Environment

ชื่อผู้แต่ง : PrathyushaRani Merla, Yiheng Liang

บทความนี้นำเสนอเกี่ยวกับการดึงข้อมูลจาก YouTube มาเพื่อทำการวิเคราะห์ว่าวิดีโอในหมวดหมู่ใดมีการเข้าชมมากที่สุด, 5 อันดับของผู้อัปโหลดที่มากที่สุด และ วิดีโอที่มีการเข้าชมมากที่สุด โดยเป็นการนำข้อมูล YouTube API เข้าสู่การ HDFS (Hadoop Distributed File System) ทำการ Map โดยการรวมค่าหลักเข้าด้วยกันและทำการ Reduce โดยการนับค่ารวมกันเพื่อนำผลสรุปมาใช้ประโยชน์ในการวิเคราะห์ทิศทางของความนิยมในปัจจุบันว่าเป็นไปในทิศทางใด และนำไปใช้ในการเลือกลงทุนการลงโฆษณาอย่างถูกต้อง

ชื่อเรื่อง : Sentiment Analysis of Facebook Data using Hadoop based Open Source Technologies

ชื่อผู้แต่ง : Sudipto Shankar Dasgupta, Swaminathan Natarajan, Kiran Kumar Kaipa, Sujay Kumar Bhattacharjee, Arun Viswanathan

บทความนี้นำเสนอเกี่ยวกับการสร้างระบบวิเคราะห์ความรู้สึกของผู้ใช้ Facebook ที่มีต่อสินค้าโดยใช้ Hadoop ในการสร้าง โดยใช้ 6 ยี่ห้อสินค้าในการคาดเดา ซึ่งนำคำศัพท์ในพจนานุกรมมาวิเคราะห์ โดยมีทั้งหมด 6,518 คำ แบ่งเป็นคำเชิงบวก 2,324 คำ และเชิงลบ 4,175 คำ มีค่าเฉลี่ยของความแม่นยำอยู่ที่ 67.6%

2.2 ทฤษฎีและแนวคิดที่ใช้ในการทำโครงการงาน

2.1.1 Big Data

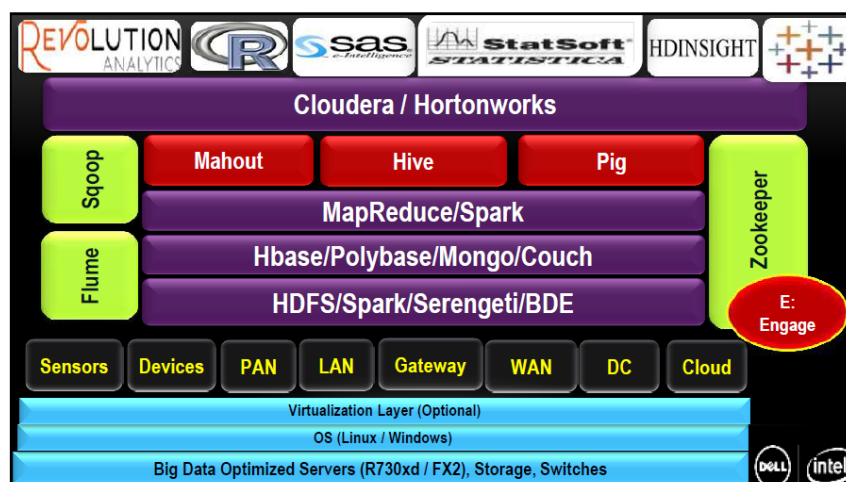
ข้อมูลขนาดใหญ่มากจนซอฟต์แวร์หรือฮาร์ดแวร์ธรรมดาไม่น่าจะสามารถที่จะจัดการหรือวิเคราะห์ได้อย่างมีประสิทธิภาพ ซึ่งมีการรวบรวมข้อมูลทั้ง Structured และ Unstructured มาทำการประมวลวิเคราะห์ข้อมูลและนำไปใช้ประโยชน์

โดยมี 5Vs ซึ่งได้แก่

- Volume (ข้อมูลขนาดใหญ่)
- Velocity (ข้อมูลมีการเกิดขึ้นอย่างรวดเร็วตลอดเวลา)
- Variety (ข้อมูลที่มีความหลากหลายในรูปแบบ)
- Veracity (ข้อมูลที่มีระดับคุณภาพปะปนกันไป)
- Variability (ข้อมูลอาจจะเกิดความเข้ากันไม่ได้ตลอดเวลา)

2.2.2 Big data platform

เป็นการนำ Module พื้นฐานที่เป็น Open Source Software ที่โดดเด่นนำมาประกอบขึ้นเป็น Platform ที่พร้อมใช้งาน โดยทั่วไปจะประกอบด้วยที่เก็บข้อมูล Big data, server, database, ตัวจัดการ Big data



รูปที่ 2-1 Big Data Platform
(ที่มา : <https://medium.com/softnix>)

2.2.3 Cludera

Cludera เป็นแพลตฟอร์มซอฟต์แวร์สำหรับการจัดการและการวิเคราะห์ข้อมูลที่มีจำนวนมหาศาล โดยพัฒนาต่อยอดจาก Apache Hadoop Platform ซึ่งเป็นซอฟต์แวร์ Open-source

ในปัจจุบัน Cludera ถือเป็น Apache Hadoop Platform หนึ่งในที่รวดเร็วที่สุด ง่ายที่สุด และปลอดภัยที่สุดในโลก ที่พัฒนาขึ้นมาเพื่อช่วยให้องค์กรธุรกิจสามารถใช้ประโยชน์จากข้อมูลจำนวนมหาศาลที่ตนเองครอบครองได้สูงสุด Cludera ทำให้องค์กรธุรกิจต่าง ๆ สามารถทุ่มเทให้กับการจัดการและวิเคราะห์ข้อมูลของตนโดยไม่ต้องสนใจเรื่องเทคโนโลยี Cludera เป็นแพลตฟอร์มที่มีประสิทธิภาพสูงสุดและคุ้มค่าที่สุดในการนำมาเพื่อใช้จัดการและวิเคราะห์ข้อมูลให้เกิดผลประโยชน์สูงสุดแก่องค์กรธุรกิจ

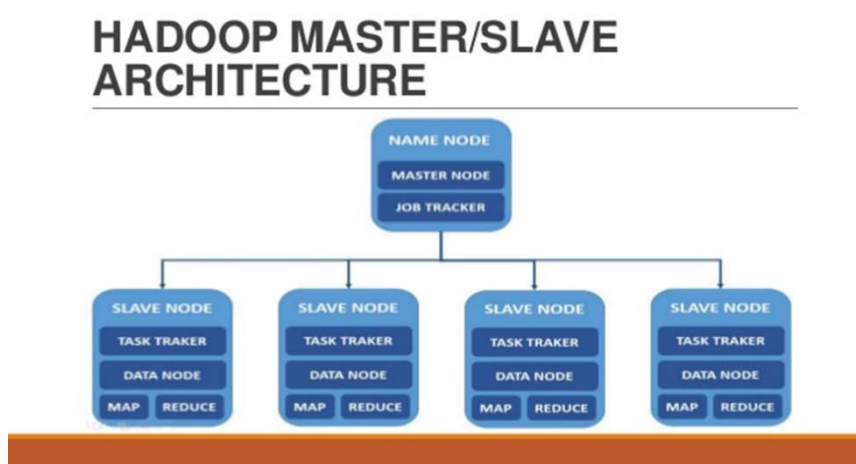
ประการแรก Cludera เป็น analytic SQL with Impala ที่เร็วที่สุด และสนับสนุน Spark Framework ที่มีประสิทธิภาพสูงสุด มี Kudu storage engine ตัวใหม่ซึ่งเป็น engine ที่ทำให้วิเคราะห์ข้อมูลที่เป็น real-time ได้รวดเร็ว นอกจากนั้น Cludera Navigator Optimizer ยังช่วยให้สามารถนำข้อมูลและภาระงานของลูกค้ามาลงในระบบและรูปแบบที่ถูกต้องเพื่อให้เกิดประสิทธิภาพสูงสุดและคุ้มค่าที่สุดทางธุรกิจ

ประการที่สอง Cludera ทำให้ Hadoop ง่ายในการจัดการทั้งบนระบบดั้งเดิมและบนระบบ Cloud สาธารณะโดยการใช้ Cludera Manager และ Cludera Director การบริการให้การช่วยเหลือสนับสนุนลูกค้า ของ Cludera ในเรื่อง Hadoop ecosystem นับว่าดีที่สุดในด้วยความเชี่ยวชาญทั้งในเชิงกว้างและเชิงลึกเกี่ยวกับ Hadoop

ประการสุดท้าย Cludera เป็นระบบที่มีความปลอดภัยอย่างสมบูรณ์และมีระบบการเข้ารหัสที่ไม่เหมือนใคร , การป้องกันการเข้าถึงข้อมูลอย่างรอบด้าน และความสามารถในการตรวจสอบและสืบย้อน (audit and lineage)

2.2.4 Hadoop

Hadoop คือ ซอฟต์แวร์ประเภท open source ที่จัดทำขึ้นเพื่อเป็นแพลตฟอร์มในการจัดเก็บข้อมูล ซึ่งมีกรอบการทำงานเพื่อใช้ในการจัดเก็บข้อมูลและประมวลผลข้อมูลที่มีขนาดใหญ่หลายๆ ที่เรียกว่า Big Data ซึ่ง Hadoop สามารถปรับขยาย ยืดหยุ่น เพื่อรองรับข้อมูลที่มีจำนวนมากมายมหาศาลได้ ทั้งนี้ก็เพราะ Hadoop มีกระบวนการประมวลผลที่แข็งแกร่งมากซึ่งเป็นผลมาจากการประมวลผลข้อมูลแบบกระจายผ่านเครื่องคอมพิวเตอร์ที่ถูกจัดอยู่ในรูปแบบ Cluster อันนำไปสู่ความสามารถในการรองรับข้อมูลที่ไม่จำกัดและมีความน่าเชื่อถือสูง



รูปที่ 2-2 Hadoop Architecture

(ที่มา : <https://www.dezyre.com/article/hadoop-ecosystem-components-and-its-architecture/114>)

โดยตามหลักการแล้วนั้น Hadoop คือ Java programming framework ที่รองรับการทำงานที่ต้องประมวลผลและเก็บข้อมูลขนาดใหญ่ เป็นส่วนหนึ่งของ Apache project โดย Hadoop ถูกออกแบบมาให้เป็น Application ที่สามารถทำงานได้บนระบบแบบ Node หรือมี Hardware จำนวนหลายๆเครื่องพร้อมกัน เพื่อรองรับข้อมูลขนาดใหญ่ ซึ่งใช้การแตก file system ออกมากระจายตาม node ให้สามารถทำงานได้รวดเร็วพร้อมทั้งส่งข้อมูลหากันระหว่าง node ทั้งหมด รวมถึงมีความสามารถในการจัดการ node ที่เสียหายได้โดยไม่ทำให้เกิดข้อมูลสูญหาย ปัจจุบัน Hadoop ถูกนำมาใช้ในงาน big data จำพวก

การคำนวณข้อมูลทางวิทยาศาสตร์เฉพาะทาง, เชิงธุรกิจ รวมถึงวางแผนการขาย และ ประมวลผลข้อมูล sensor จำนวนมาก หรือ internet of things (IOT)

ส่วนประกอบของ Hadoop นั้น เนื่องจากเป็น Software framework ทำให้ Hadoop ประกอบไปด้วย module จำนวนมาก ซึ่งขั้นต่ำที่จะต้องมียังคือ

- Hadoop Common: เป็น library ที่จำเป็นในการใช้งานเปรียบได้กับ kernel
- Hadoop Distributed File System (HDFS): สำหรับเก็บข้อมูลภายใต้ server node จำนวนมาก ซึ่งต้องอาศัย bandwidth ที่สูงมากในการทำงาน
- Hadoop Yet Another Resource Negotiator (YARN): ทำหน้าที่จัดการทรัพยากรที่มีอยู่ รวมถึงรอบการทำงานของ application
- Hadoop MapReduce: เป็นรูปแบบ program ที่จัดการกับข้อมูลขนาดใหญ่เพื่อสร้างความสัมพันธ์ระหว่างกัน

โดยในส่วนของการทำงาน โครงสร้างของ Hadoop จะเป็นแบบ Cluster โดยแบ่งออกเป็น 2 ส่วนคือ Master server และ Slave server ซึ่งมีการทำงานหลักๆอยู่ 3 ส่วนด้วยกัน ได้แก่

- Cluster management
- Storage
- Processing

2..2.5 Node

Namenode

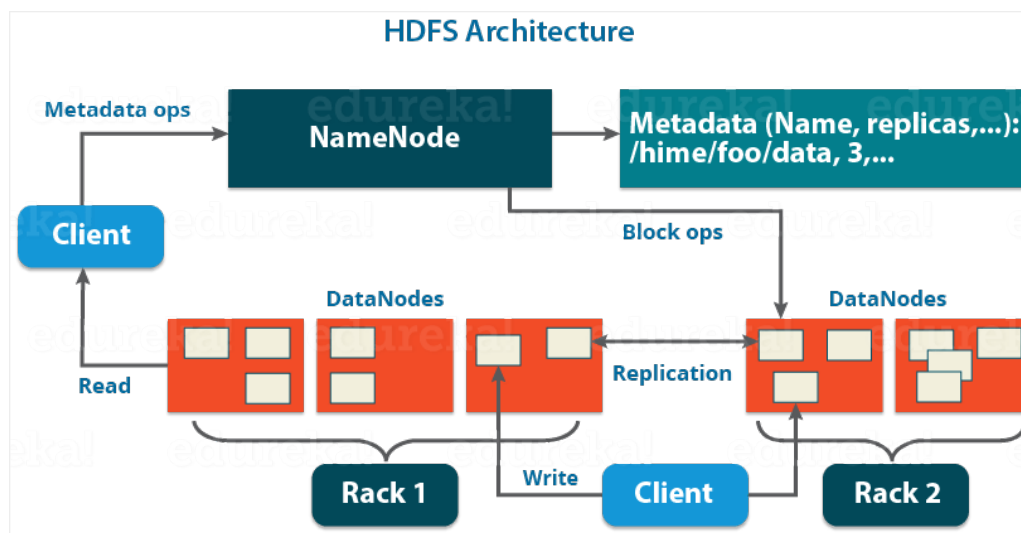
- เป็นโหนดหลัก ไม่ได้ใช้ในการเก็บข้อมูลจริงๆ
- ใช้ในการเก็บ metadata
- ต้องการพื้นที่น้อยแต่ต้องการการประมวลผลเยอะ

Datanode

- ข้อมูลจะถูกเก็บในโหนดนี้
- เป็นเครื่องคอมพิวเตอร์ทั่วไป
- มีการกระจายมาที่โหนดเหล่านี้ เป็นเหตุผลที่ทำให้ Hadoop มีประสิทธิภาพสูง

2.2.6 HDFS (Hadoop Distributed File System)

คือระบบจัดเก็บข้อมูลหลัก ที่ใช้ในซอฟต์แวร์ Hadoop ซึ่ง HDFS จะสร้างแบบจำลองเป็นบล็อกข้อมูลบนคลัสเตอร์เพื่อให้เกิดความน่าเชื่อถือและการคำนวณผลที่รวดเร็ว ซึ่งถือเป็นตัวที่ใช้ในการเก็บข้อมูลต่างๆ ที่มีความแตกต่างกัน และนอกจากนี้ยังช่วยในการเก็บข้อมูลไว้ในหลายๆ โหนดและจัดการ log ไฟล์ในการเก็บข้อมูล (metadata)



รูปที่ 2-3 HDFS Architecture

(ที่มา : <https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/>)

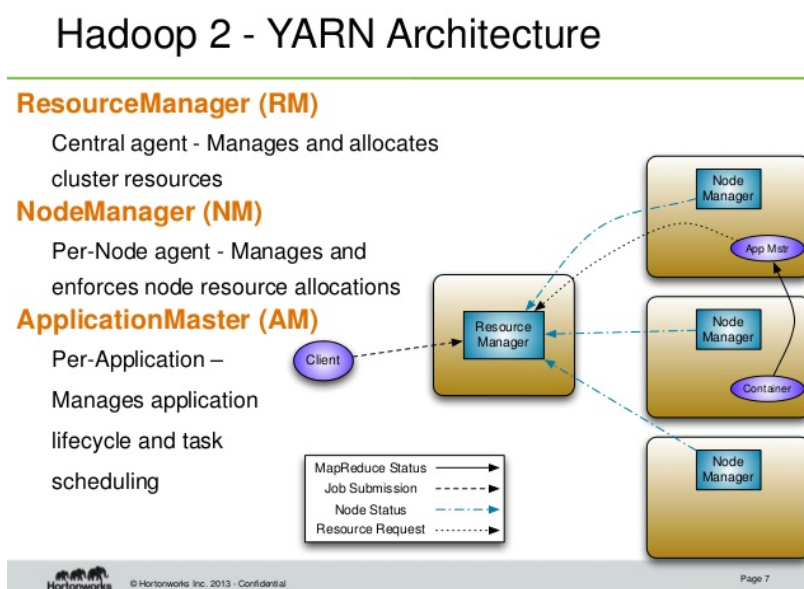
2.2.7 Yarn (Yet Another Resource Negotiator)

เป็นส่วนช่วยในการจัดการ resource ภายใน hadoop cluster ซึ่งใน version นี้เอง client ไม่จำเป็นต้องเรียกใช้งาน framework เช่น MapReduce หรือ Spark เพื่อติดต่อและจัดการ resource เองแต่จะผ่าน YARN และตัว YARN จะไปทำงานร่วมกับ resource layer อย่าง HDFS เอง

ภายใน YARN จะประกอบไปด้วยส่วนประกอบหลัก 2 ส่วนคือ resource manager และ node manager โดยที่ resource manager จะทำหน้าที่คอยดูแล resource ของทั้ง cluster ซึ่งจะมีแค่ 1 ตัวต่อ cluster เท่านั้น ส่วน node manager จะทำหน้าที่ดูแลจัดการและ monitor resource ในแต่ละ node หรือ

container รวมถึง launch โปรแกรมที่จะมาใช้ resource ใน node นั้นๆและติดต่อไปยัง resource manager ที่ดูแลระดับ cluster โดย node manager นี้จะถูกติดตั้งตามจำนวน node ใน cluster

ในการจะ launch โปรแกรม (application master) นั้น ส่วน client ของ user ต้องติดต่อกับทาง resource manager เพื่อให้ทาง resource manager จัดการหา node manager ที่มีความที่จะรับ application master ไป launch ที่ node ได้ โดย application master อาจจะเป็น application ที่ run จบแล้วส่ง result ของการ process กลับไปยัง client ได้เลยหรืออาจจะเป็น application ที่ต้องการการ process ที่ซับซ้อนและต้องการ resource จาก resource manager เพิ่มระหว่างการ process



รูปที่ 2-4 YARN Architecture

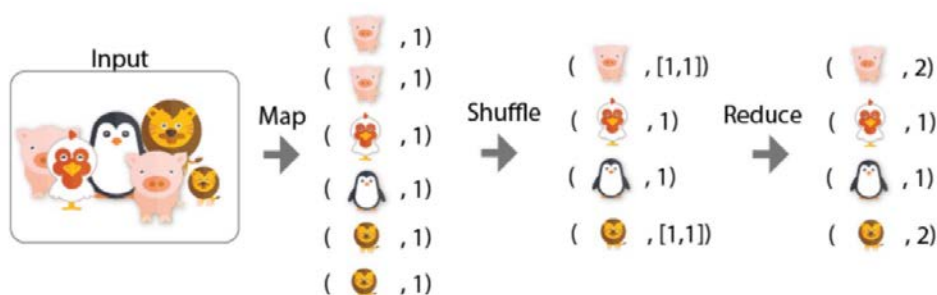
(ที่มา : <https://www.slideshare.net/hortonworks/apache-hadoop-yarn-understanding-the-data-operating-system-of-hadoop>)

2.2.8 MapReduce

เป็นแกนในการประมวลผลของ Hadoop โดยคอยให้ logic ในการประมวลผลและช่วยในการประมวลผลดาต้าเซตใหญ่ๆโดยใช้การกระจายและอัลกอริทึมแบบคู่ขนาน โดยที่

Map() จะทำหน้าที่เป็นฟิเตอร์ จับข้อมูลรวมกลุ่ม และเรียงข้อมูล

Reduce() ทำหน้าที่รวมผลและสรุปผลจากผลลัพธ์ที่ได้จากการ Map()



รูปที่ 2-5 Map Reduce

(ที่มา : http://www.howtoautomate.in.th/the-big-data-world-explanation-part-one/2018-05-29-21_58_49-mapreduce_bhuridech-sudsee-medium/)

2.2.9 Hive

กระบวนการทำงานของ Hive คือการทำหน้าที่ส่งคำสั่งประมวลผลข้อมูลผ่านเข้าสู่ Hadoop ซึ่งการประมวลผลใดๆกับข้อมูล ของ Hadoop ยังคงทำผ่าน MapReduce ตามปกติ

ดังนั้นบทบาทของ Hive จะเป็นตัวกลางที่มี Engine ในการแปลงภาษา HQL หรือ SQL ในแบบของ Hive เป็น MapReduce เพื่อส่งไปทำงานที่ Hadoop Engine โดยกระบวนการทำงานของ Hive Engine จำเป็นต้องมี Metadata ที่เป็นข้อมูลอธิบายรายละเอียดของโครงสร้างเหมือนตารางข้อมูล ประเภทข้อมูลที่อ้างอิงเป็นคอลัมน์ข้อมูล ซึ่งภาพของการจัดการกับ Metadata คือ การสร้าง Table หรือ Create Table แบบภาษา SQL

การที่มี Metadata ของ Table ก็เพื่อให้ Hive Engine และ Hadoop Engine ทำงานร่วมกัน โดยหลักการทำงานคือการแปลง HQL เป็น MapReduce ในการประมวลผลข้อมูล หรือ แปลงผลลัพธ์การประมวลผลข้อมูลจาก MapReduce

ที่อ่านจากไฟล์คืนกลับเป็นข้อมูลแบบมีโครงสร้างแบบตารางข้อมูลที่เอื้อต่อการมี API

ของแต่ละภาษาโปรแกรมมิ่งมาต่อยอดนำไปสู่การทำวิเคราะห์ข้อมูล ประมวลผลข้อมูลได้อย่างง่าย

ดังนั้นส่วนประกอบที่สำคัญของ Hive ที่ Hive จะสร้าง MapReduce Job ตอนที่เรทำการบันทึกข้อมูล หรือ เรียกดูข้อมูล คือตัว Metastore

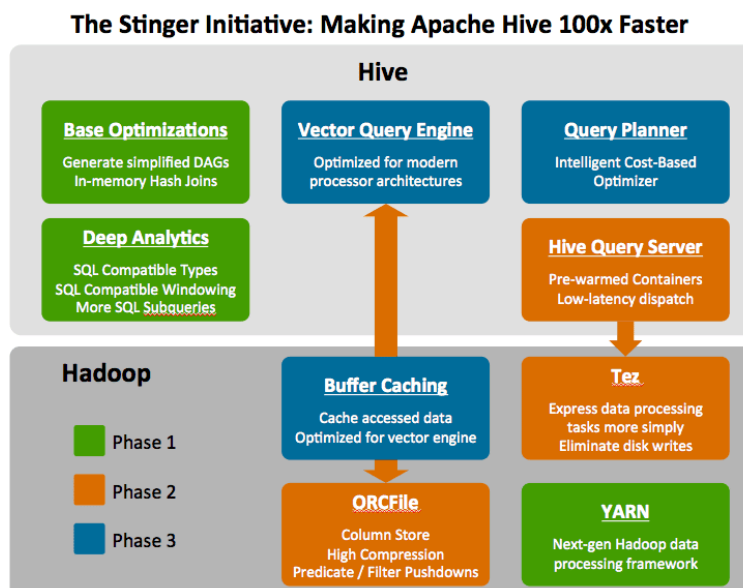
ดังนั้น Big Data Processing ด้วย Hive จะเป็นลักษณะงานที่เป็นแบบ Batch Processing

เพื่อวิเคราะห์ข้อมูลสำหรับหาคำตอบหนึ่งๆหรือค่าสรุปใดๆจากข้อมูลเป็นปริมาณมาก หรือ

งานแบบเรียกข้อมูลมาแสดงผลที่จากปริมาณข้อมูลมากๆแต่การใช้งานยอมรับได้เช่น

ค้นหาและแสดงประวัติข้อมูล การแสดงรายงานแบบต่างๆ ดังนั้นลักษณะงานหรือแอปพลิเคชันที่เหมาะสมสำหรับ Hive จะเป็นตัวอย่างดังนี้

- Data Mining
- Log Processing
- Document Indexing
- Customer Facing Business Intelligence
- Predictive Modelling
- Hypothesis Testing

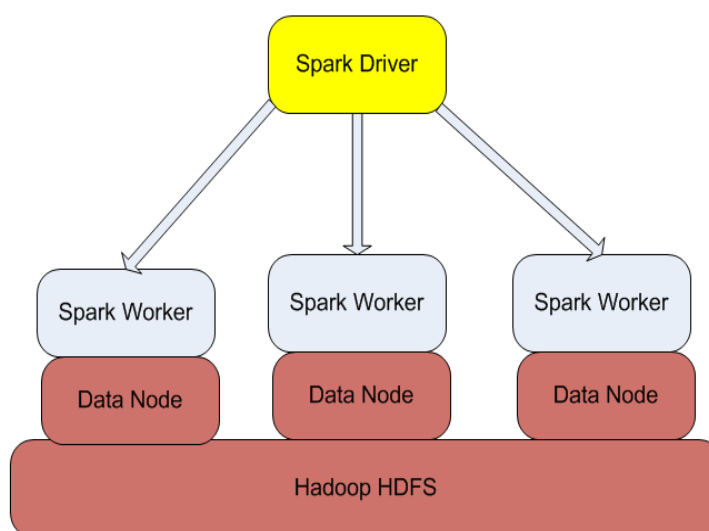


รูปที่ 2-6 Making Apache Hive 100x faster

(ที่มา : <https://hortonworks.com/blog/100x-faster-hive/>)

2.2.10 Spark

Apache Spark เป็น Processing Engine ที่เป็น Open Source ซึ่งใช้ทำงานในด้าน Big Data และ Data Science โดยมีหลักการทำงานเหมือนกับ MapReduce และมีการประมวลผลบนหน่วยความจำเป็นหลัก ทำให้ทำงานได้อย่างรวดเร็ว และยังเป็น Processing Engine ที่เป็น Open Source สามารถใช้ทำงานในด้าน Big Data และ Data Science มีความเร็วในการทำงานที่มากกว่า MapReduce มากและใช้ในการประมวลผลแบบเรียลไทม์สำหรับระบบแบบกระจายข้อมูล



รูปที่ 2-7 Spark Architecture

(ที่มา : <https://bighadoop.wordpress.com/2014/04/03/apache-spark-a-fast-big-data-analytics-engine/>)

2.2.11 ZooKeeper

- เป็นศูนย์กลางบริการในการจัดการกับข้อมูล
- เป็นตัวจัดการการกระจาย services ภายในระบบ
- เป็นตัวควบคุมการทำงานต่างๆ ในระบบให้เป็นไปอย่างราบรื่นและจัดการกับ Bug ที่เกิดขึ้น
- เป็นตัวประสานงานภายใน Hadoop

2.2.12 RapidMiner

การวิเคราะห์ข้อมูลด้วย Data Mining กำลังเป็นที่นิยมไปทั่วโลกด้วยแรงขับเคลื่อนอย่างหนึ่งคือการมีซอฟต์แวร์ที่ช่วยให้ทำการวิเคราะห์ได้ง่ายขึ้น เหมาะกับการวิเคราะห์ข้อมูลที่มีขนาดใหญ่และจำนวนมาก เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล

สำหรับการนำมาใช้เพื่อการบริหารและการตัดสินใจขององค์กร ซึ่ง Data Mining Tools ที่ได้รับความเชื่อมั่นในประสิทธิภาพอย่างมากอันหนึ่งคือ RapidMiner ซึ่งเป็น open source software ทำให้สามารถประหยัดค่าใช้จ่าย กรณีต้องการใช้งานเพื่อทดสอบการวิเคราะห์ข้อมูล การทำความสะอาดข้อมูล (Data cleansing) และการรายงานข้อมูลในรูปแบบต่างๆ เช่น ตาราง กราฟ เป็นต้น



(ที่มา : <https://rapidminer.com/>)

2.2.13 Radoop

การที่จะเชื่อมต่อระหว่าง Rapidminer เข้ากับ Cloudera ได้นั้นย่อมต้องมีตัวกลาง ซึ่ง Radoop นั้นเปรียบได้ดังผู้ที่เข้ามาเพื่อลดกำแพงในการใช้ระบบ Hadoop ลง Radoop นั้นเหมาะกับทั้งผู้ที่เริ่มสนใจและศึกษา Big data ที่ต้องการเข้าถึง, วิเคราะห์ และ ทำให้ Big data เป็นรูปเป็นร่างได้ นอกจากนั้น Radoop ยังเหมาะกับผู้เชี่ยวชาญในด้านนี้อยู่แล้วอีกด้วย โดย Radoop สามารถเขียนสคริป, ทำการประมวลผล และทำการคาดการณ์ระดับสูงได้



รูปที่ 2-9 Radoop

(ที่มา : <https://dataconomy.com/2014/04/radoop-making-big-data-analytics-easy/>)

บทที่ 3

วิธีการดำเนินโครงการ

3.1 วิธีดำเนินการวิจัยและวิธีการประเมินผล

วิธีการดำเนินการวิจัย

- ศึกษาข้อมูลเกี่ยวกับ Big Data, Hadoop และ RapidMiner
- วิเคราะห์และออกแบบระบบ
- เตรียมเครื่องมือพัฒนา
- สร้างและพัฒนา
- ทดสอบการใช้งาน
- ประเมินผลการใช้งาน
- สรุปผลการดำเนินงาน
- จัดทำรูปเล่มวิจัย

วิธีการประเมินผล

- ประเมินและเปรียบเทียบสมรรถนะในการวิเคราะห์ข้อมูล Big Data Platform กับเครื่องคอมพิวเตอร์ทั่วไป

3.2 แผนการดำเนินงานตลอดโครงการ

	แผนการดำเนินงาน	2561					2562				
		ส.ค	ก.ย	ต.ค	พ.ย	ธ.ค	ม.ค	ก.พ	มี.ค	เม.ย	พ.ค
1	ศึกษาข้อมูลและวิธีการพัฒนา										
2	วิเคราะห์และออกแบบระบบ										
3	เตรียมเครื่องมือการพัฒนา										
4	สร้างและพัฒนา										
5	ทดสอบการใช้งาน										
6	ประเมินผลการใช้งาน										
7	สรุปผลการดำเนินงาน										
8	จัดทำรูปเล่มงานวิจัย										

ตารางที่ 3.1 ตารางการดำเนินงาน

3.3 อุปกรณ์และเครื่องมือที่ใช้ในโครงการ

3.3.1 ฮาร์ดแวร์ (Hardware)

PC 3 เครื่อง

PC (1)

- Processor : Intel Core i7
- Memory : 16 GB
- Storage : 1TB

PC (2)

- Processor : Intel Core i5
- Memory : 12 GB
- Storage : 1TB

PC (2)

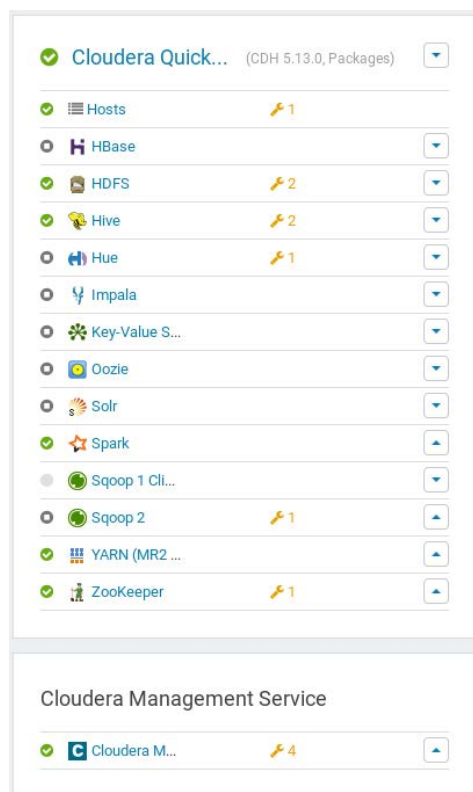
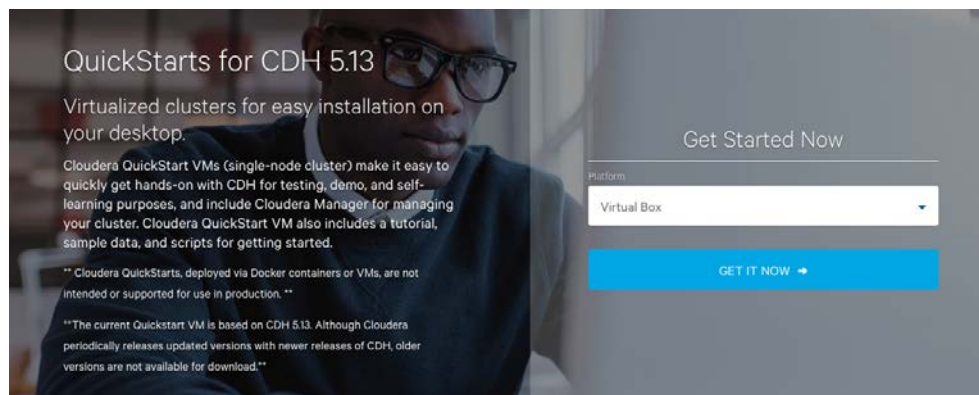
- Processor : Intel Core i5
- Memory : 8 GB
- Storage : 1TB

3.3.2 ซอฟต์แวร์ (Software)

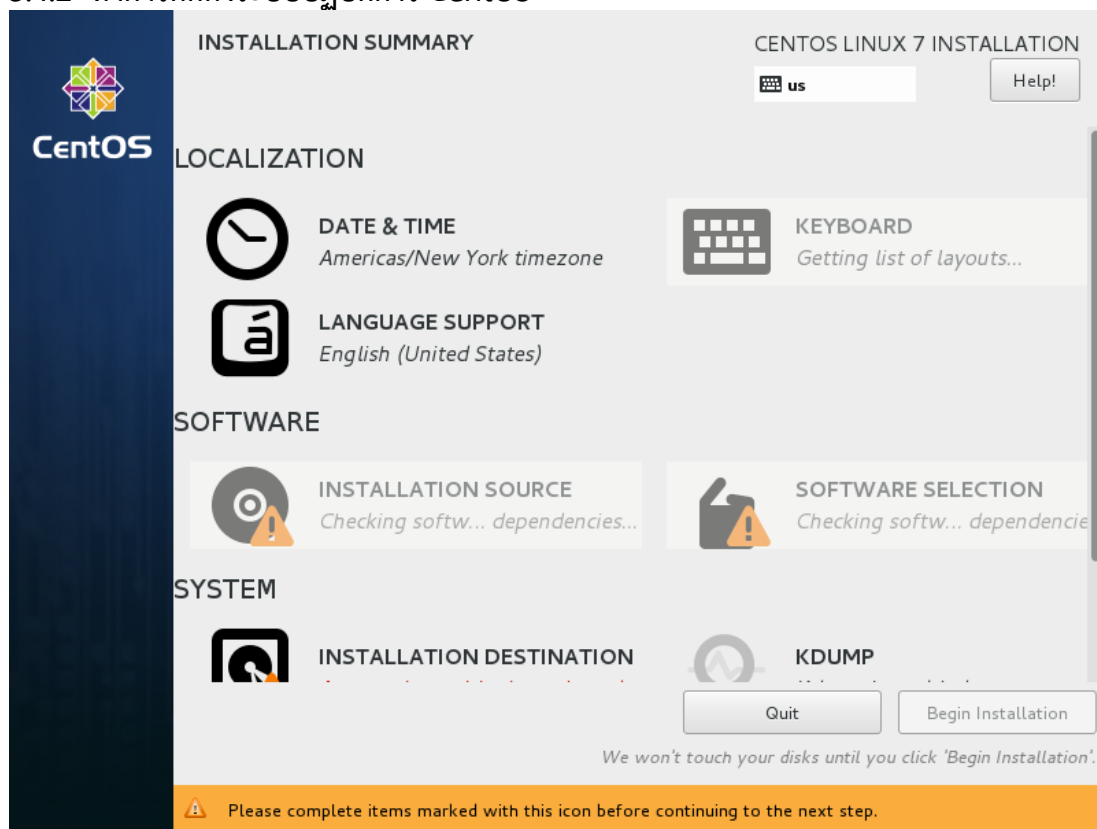
- CentOS 7.0
- Cloudera 6.0.1
- Hadoop
- Rapid Miner
- Radoop

3.4 ขั้นตอนการดำเนินงาน

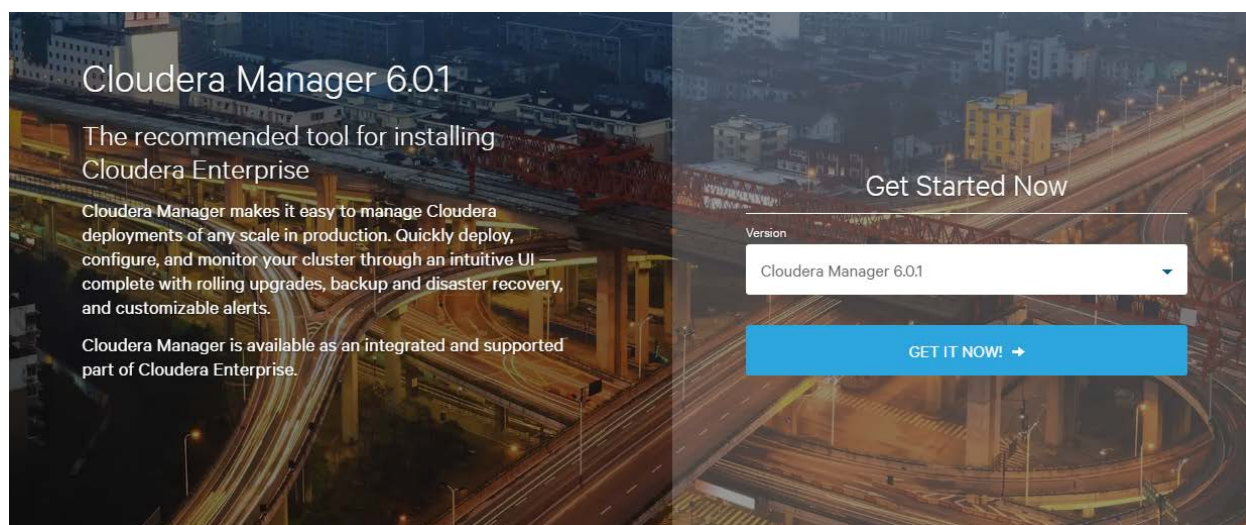
3.4.1 ศึกษาการทำงานเบื้องต้นของ Cloudera โดยการทดลองใช้งาน Cloudera Quickstart บน Virtual Machine



3.4.2 ทำการติดตั้งระบบปฏิบัติการ CentOS



3.4.3 ติดตั้ง Cloudera Manager



3.4.4 ทำการเพิ่ม Namenode และ Datanode

Cloudera **Manager**

Add New Hosts to Cluster

Specify Hosts

Specify hosts for your CDH cluster installation. Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Hostname:

Hint: Search for hostnames or IP addresses using [patterns](#)

SSH Port:

3 hosts scanned, 3 running SSH.
Click the first checkbox, hold down the Shift key and click the last checkbox to select a range.

☒	Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
☐	master	master.swu.ac.th	10.1.144.121	Yes	Host was successfully scanned.
☐	node1	node1.swu.ac.th	10.1.144.173	Yes	Host was successfully scanned.
☐	node2	node2.swu.ac.th	10.1.144.116	Yes	Host was successfully scanned.

Displaying 1 - 3 of 3

3.4.5 ทำการติดตั้ง Cloudera Agent ใน namenode และ Datanode

Cloudera **Manager**

Add Cluster - Installation

Install Agents

Installation in progress.

2 of 3 host(s) completed successfully. [Abort Installation](#)

Hostname	IP Address	Progress	Status
master.swu.ac.th	10.1.144.121		✓ Installation completed successfully. Details
node01.swu.ac.th	10.1.134.137		✓ Installation completed successfully. Details
node02.swu.ac.th	10.1.144.116		○ Installing cloudera-manager-agent package... Details

3.4.6 ทำการ Download และ Distribute Parcels ให้แก่ Node ต่างๆ

Cluster Installation

Installing Selected Parcels

The selected parcels are being downloaded and installed on all the hosts in the cluster.

Parcel	Downloaded	Distributed	Unpacked	Activated
CDH 5.10.0-1.cdh5.10.0.p0.41	100%	4/4 (93.4 MiB/s)	4/4	4/4
ORACLE_INSTANT_CLIENT 11.2-1.oracleinst...	100%	4/4 (54.2 MiB/s)	4/4	4/4

3.4.7 ทำการเลือก service ต่างๆ ที่ต้องการใช้งาน

Setup Database

Review Changes

Command Details

Summary

HDFS

NameNode × 1 New
master.swu.ac.th

SecondaryNameNode × 1 New
master.swu.ac.th

Balancer × 1 New
master.swu.ac.th

HttpFS
Select hosts

NFS Gateway
Select hosts

DataNode × 2 New
node[01-02].swu.ac.th

Hive

Gateway × 3 New
master.swu.ac.th; node[01-02]...

Hive Metastore Server × 1 New
master.swu.ac.th

WebHcat Server
Select hosts

HiveServer2 × 1 New
master.swu.ac.th

Hue

Hue Server × 1 New
master.swu.ac.th

Load Balancer × 1 New
master.swu.ac.th

Cloudera Management Service

Service Monitor × 1 New
master.swu.ac.th

Activity Monitor
Select a host

Host Monitor × 1 New
master.swu.ac.th

Event Server × 1 New
master.swu.ac.th

Alert Publisher × 1 New
master.swu.ac.th

Oozie

Oozie Server × 1 New
master.swu.ac.th

Spark

History Server × 1 New
master.swu.ac.th

Gateway × 3 New
master.swu.ac.th; node[01-02]...

YARN (MR2 Included)

Summary

master.swu.ac.th; node[01-02]...

master.swu.ac.th

Select hosts

master.swu.ac.th

Hue

Hue Server × 1 New
master.swu.ac.th

Load Balancer × 1 New
master.swu.ac.th

Cloudera Management Service

Service Monitor × 1 New
master.swu.ac.th

Activity Monitor
Select a host

Host Monitor × 1 New
master.swu.ac.th

Event Server × 1 New
master.swu.ac.th

Alert Publisher × 1 New
master.swu.ac.th

Oozie

Oozie Server × 1 New
master.swu.ac.th

Spark

History Server × 1 New
master.swu.ac.th

Gateway × 3 New
master.swu.ac.th; node[01-02]...

YARN (MR2 Included)

ResourceManager × 1 New
master.swu.ac.th

JobHistory Server × 1 New
master.swu.ac.th

NodeManager × 2 New
Same As DataNode

ZooKeeper

Server × 1 New
master.swu.ac.th

Back

Continue

3.4.8 เปิดใช้งาน node ต่างๆ

Cloudera Manager Clusters Hosts Diagnostics Audits Charts Administration

Search  Support admin

All Hosts Configuration Add New Hosts to Cluster Review Upgrade Status Inspect All Hosts

Search

Filters

- STATUS
 - Concerning Health 3
- CLUSTERS
- CORES
- COMMISSION STATE

Actions for Selected Columns: 10 Selected

Status	Name	IP	Roles	Commission State	Last Heartbeat	Load Average	Disk Usage	Physical Memory	Swap Space
	master.svu.ac.th	10.1.144.121	> 18 Role(s)	Commissioned	7.24s ago	0.94 1.03 0.78	16.8 GB / 243.3 GB	8 GB / 15.5 GB	0 B / 7.8 GB
	node1.svu.ac.th	10.1.144.173	> 4 Role(s)	Commissioned	4.75s ago	0.07 0.15 0.22	11 GB / 725.4 GB	2.8 GB / 11.5 GB	0 B / 5.9 GB
	node2.svu.ac.th	10.1.144.116	> 4 Role(s)	Commissioned	6.85s ago	0.20 0.15 0.10	13 GB / 373.6 GB	2.4 GB / 7.6 GB	0 B / 7.7 GB

3.4.9 ติดตั้ง Service ให้ node ต่างๆ

Add HBase Service to Namenode Test

First Run Command

Namenode Test (CDH 5.0.0, Packages) ▼

1 Hosts	1	
HBase		
HDFS	1	2
Hive		
Impala		
Spark		
YARN (MR2 Inc...)		
ZooKeeper	1	

Charts

Cluster CPU

■ Namenode Test, Host CPU Usage Across Hosts 11.4%

Cloudera Management Service

Clou **Namenode Test** (CDH 5.0.0, Packages) ▼

1 Hosts	1	
HBase		
HDFS	2	
Hive		
Impala		
Spark		
YARN (MR2 Inc...)		
ZooKeeper	1	

Cluster Disk IO

■ Total Disk Byte... 6.8K/s ■ Total Bytes Written... 0.93b/s

Completed Impala Queries

queries / second

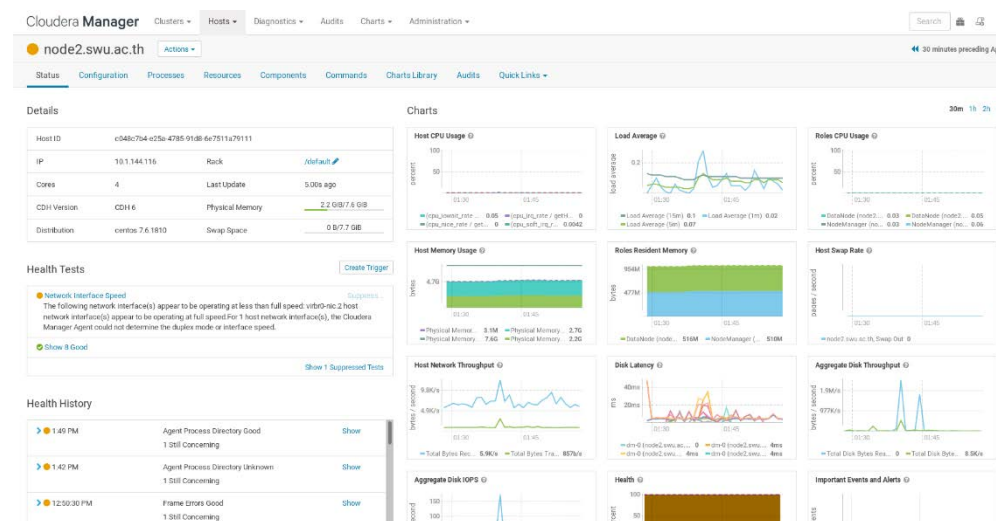
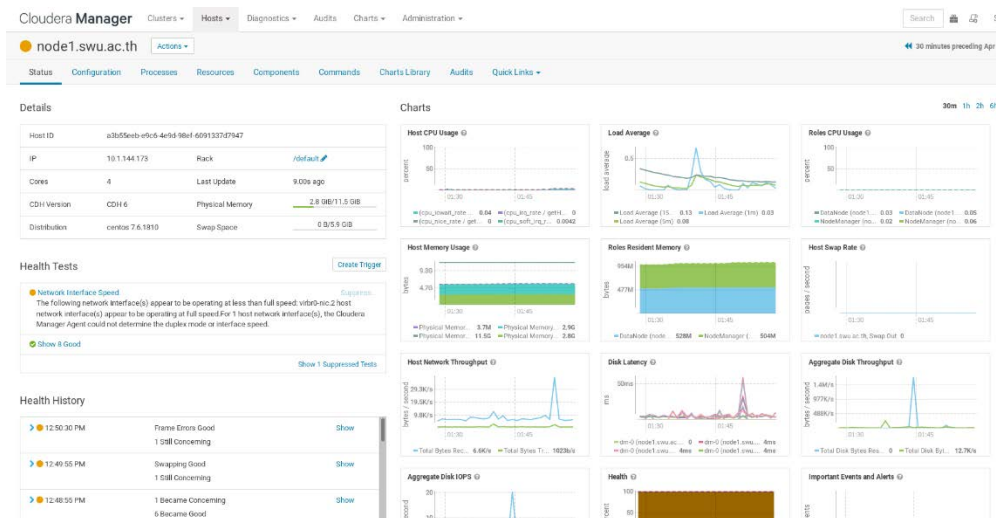
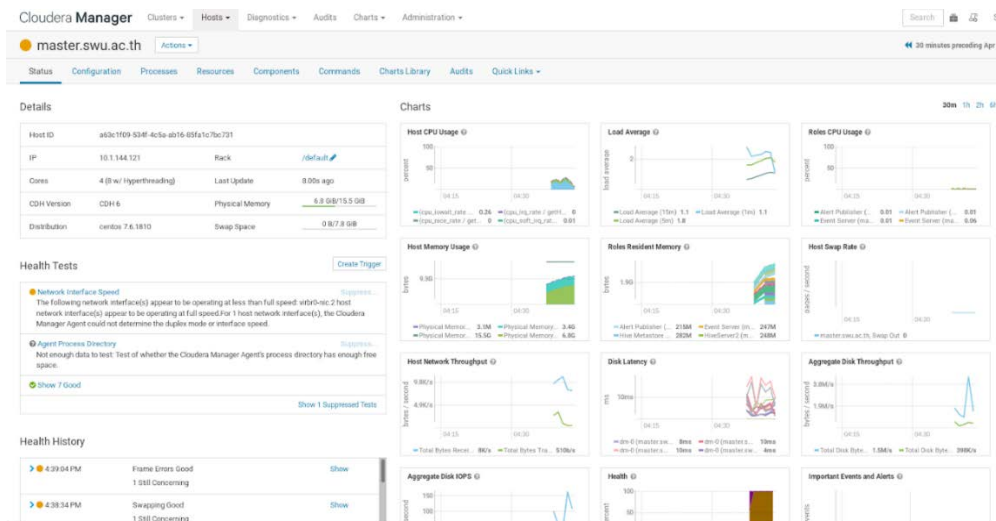
NO DATA

30m 1h 2h 6h 12h 1d 7d 30d

■ Completed ■ In Progress

3.4.10 เปิดใช้งาน Cloudera Manager และ Service ต่างๆ

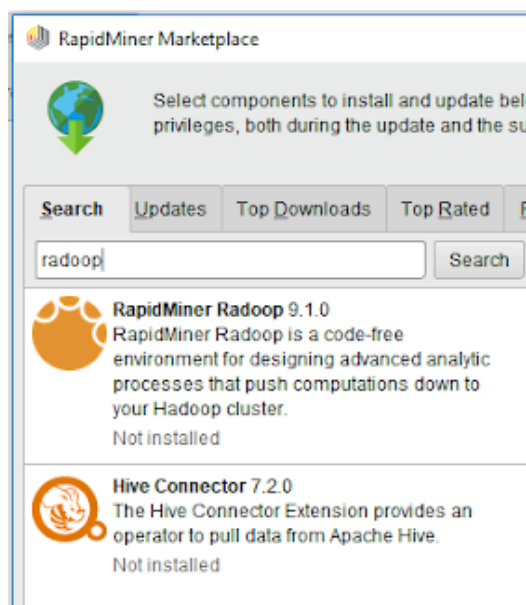
3.4.11 เช็สถานะการทำงานของ node ต่างๆ



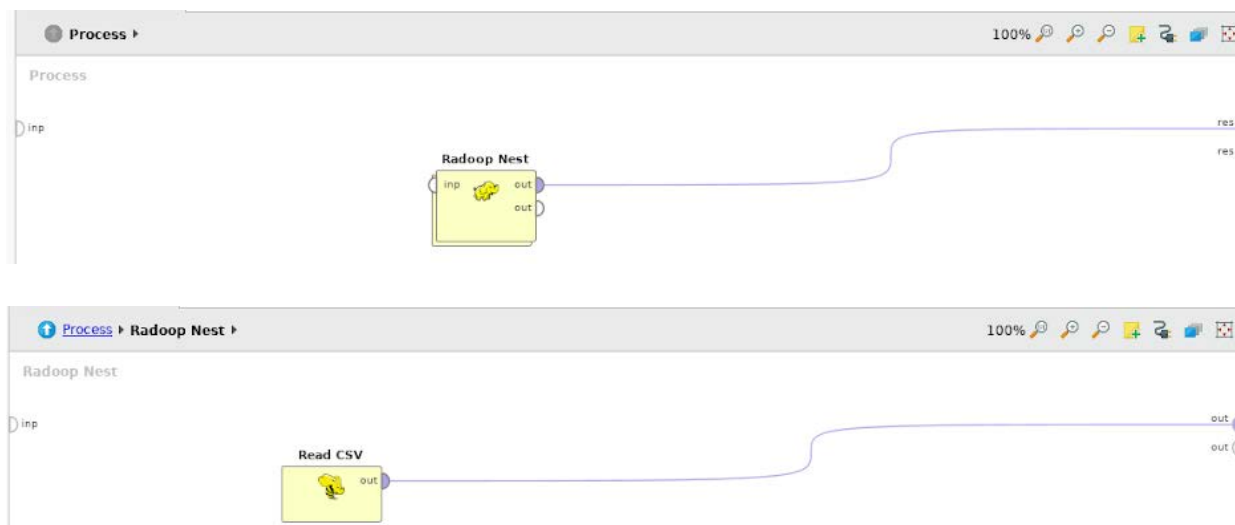
3.4.12 ติดตั้ง RapidMiner



3.4.13 ติดตั้ง RapidMiner Radoop และ ทำการเชื่อมต่อระหว่าง rapidMiner Radoop กับ Cluster



3.4.14 ทดสอบการอ่านไฟล์



3.4.15 ผลลัพธ์จากการทดสอบการอ่านไฟล์

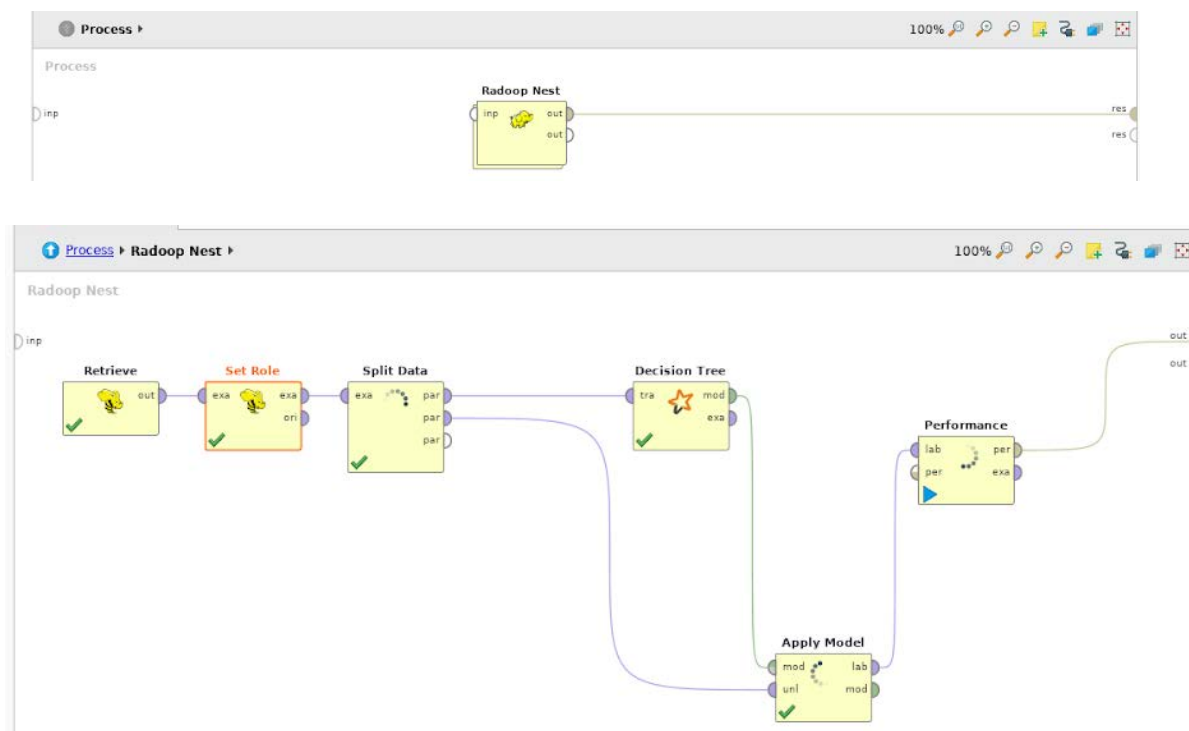
Open in [Turbo Prep](#) [Auto Model](#)

Filter (151 / 151 examples): all

Row No.	att0	att1	att2	att3	att4
1	5.100	3.500	1.400	0.200	Iris-setosa
2	4.900	3	1.400	0.200	Iris-setosa
3	4.700	3.200	1.300	0.200	Iris-setosa
4	4.600	3.100	1.500	0.200	Iris-setosa
5	5	3.600	1.400	0.200	Iris-setosa
6	5.400	3.900	1.700	0.400	Iris-setosa
7	4.600	3.400	1.400	0.300	Iris-setosa
8	5	3.400	1.500	0.200	Iris-setosa
9	4.400	2.900	1.400	0.200	Iris-setosa
10	4.900	3.100	1.500	0.100	Iris-setosa
11	5.400	3.700	1.500	0.200	Iris-setosa
12	4.800	3.400	1.600	0.200	Iris-setosa
13	4.800	3	1.400	0.100	Iris-setosa
14	4.300	3	1.100	0.100	Iris-setosa
15	5.800	4	1.200	0.200	Iris-setosa
16	5.700	4.400	1.500	0.400	Iris-setosa
17	5.400	3.900	1.300	0.400	Iris-setosa
18	5.100	3.500	1.400	0.300	Iris-setosa
19	5.700	3.800	1.700	0.300	Iris-setosa
20	5.100	3.800	1.500	0.300	Iris-setosa
21	5.400	3.400	1.700	0.200	Iris-setosa
22	5.100	3.700	1.500	0.400	Iris-setosa
23	4.600	3.600	1	0.200	Iris-setosa
24	5.100	3.300	1.700	0.500	Iris-setosa

ExampleSet (151 examples, 0 special attributes, 5 regular attributes)

3.4.16 ทำการทดสอบระบบโดยการทำ Decision Tree



3.4.17 ผลลัพธ์จากการทำ Decision Tree

accuracy: 90.00%				
	true Iris-versicolor	true Iris-virginica	true Iris-setosa	class precision
pred. Iris-versicolor	12	0	2	85.71%
pred. Iris-virginica	2	13	0	86.67%
pred. Iris-setosa	0	0	11	100.00%
class recall	85.71%	100.00%	84.62%	

3.4.18 ตรวจสอบการทำงานของ Cluster

Applications Places Firefox

All Applications - Mozilla Firefox

Internet Passport: ระบุบัญชี x Status - Home - Cloudera x Cloudera Manager x Problem loading page x New Tab x All Applications x

master:8088/cluster

Scheduler Metrics

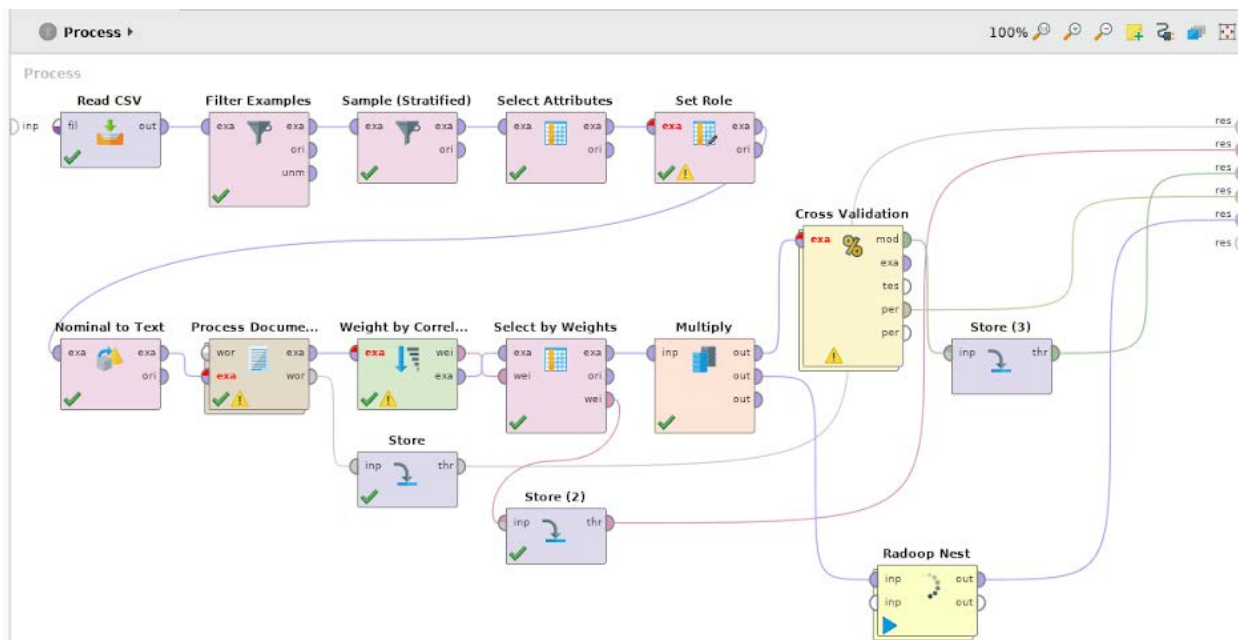
Scheduler Type: Fair Scheduler Scheduling Resource Type: [memory:mb (unit=M), vcores] Minimum Allocation: <memory:1024, vCores:1> Maximum Allocation: <memory:1762, vCores:4> Maximum Cluster App: 0

Show 20 entries

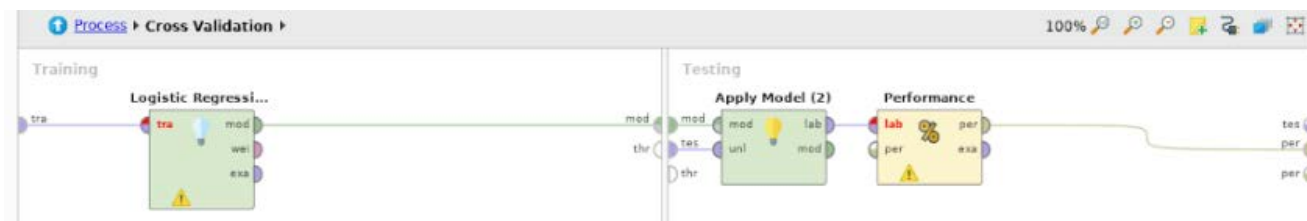
ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster
application_1556257838028_0014	hive	Performance	MAPREDUCE	root.users.hive	0	Fri Apr 26 14:10:12 +0700 2019	Fri Apr 26 14:10:29 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0013	root	Decision Tree	SPARK	root.users.root	0	Fri Apr 26 14:09:49 +0700 2019	Fri Apr 26 14:10:08 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0012	hive	Decision Tree	MAPREDUCE	root.users.hive	0	Fri Apr 26 14:09:34 +0700 2019	Fri Apr 26 14:09:47 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0011	hive	Split Data	MAPREDUCE	root.users.hive	0	Fri Apr 26 14:09:18 +0700 2019	Fri Apr 26 14:09:29 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0010	root	Spark job	SPARK	root.users.root	0	Fri Apr 26 14:08:21 +0700 2019	Fri Apr 26 14:08:30 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0009	root	Decision Tree	SPARK	root.users.root	0	Fri Apr 26 14:06:08 +0700 2019	Fri Apr 26 14:06:24 +0700 2019	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0008	hive	Decision Tree	MAPREDUCE	root.users.hive	0	Fri Apr 26 14:05:53 +0700 2019	Fri Apr 26 14:06:05 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0007	hive	Split Data	MAPREDUCE	root.users.hive	0	Fri Apr 26 14:05:36 +0700 2019	Fri Apr 26 14:05:48 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0006	root	Decision Tree	SPARK	root.users.root	0	Fri Apr 26 14:04:24 +0700 2019	Fri Apr 26 14:04:43 +0700 2019	FINISHED	FAILED	N/A	N/A	N/A	N/A	N/A	0.0	0.0
application_1556257838028_0005	hive	Decision	MAPREDUCE	root.users.hive	0	Fri Apr 26 14:04:12 +0700 2019	Fri Apr 26 14:04:24 +0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0

All Applications - Mozilla Firefox Pictures master@master:/home/master/Dow... Facebook - Mozilla Firefox (Private ... /Local_Repository/processes/firstD... 1/4

3.4.19 ทำการ Classification Model



3.4.20 ทำ Classification Model ในส่วน Cross Validation และ Process Document form Data



3.4.21 ข้อมูลที่ได้จากการทำ Classification Model

Open in Turbo Prep Auto Model Filter (200 / 200 examples): all

Row No.	existence	account	acid	action	adapt	addthi	adrian	aid	air	al	alan	algor	alixtl
1	Y	0	0	0	0	0	0	0	0.433	0	0	0	0
2	Y	0	0	0	0	0	0	0	0	0	0	0	0
3	Y	0	0	0	0	0	0	0	0	0	0	0	0
4	Y	0	0	0	0	0	0	0	0	0	0	0	0
5	Y	0	0	0	0	0	0	0	0	0	0	0	0
6	Y	0	0	0	0	0	0	0	0	0	0	0	0
7	Y	0	0	0	0	0	0	0	0	0	0	0	0
8	Y	0	0	0	0	0	0	0	0	0	0	0	0
9	N	0	0	0	0	0	0	0	0	0	0	0	0
10	Y	0	0	0	0	0	0	0	0	0	0	0	0
11	Y	0	0	0.306	0	0	0	0	0	0	0	0	0
12	Y	0	0	0.306	0	0	0	0	0	0	0	0	0
13	Y	0	0	0	0	0	0	0	0.236	0	0	0	0
14	Y	0	0	0	0	0	0	0	0	0	0	0	0
15	Y	0	0	0	0.246	0	0	0	0	0	0	0	0
16	Y	0	0	0	0	0	0	0	0	0	0	0	0
17	Y	0.275	0	0	0	0	0	0	0	0	0	0	0
18	Y	0	0	0	0	0	0	0	0	0	0	0	0
19	Y	0	0	0	0	0	0	0	0	0	0	0	0
20	Y	0	0	0	0	0	0	0	0	0	0	0	0
21	Y	0	0	0	0	0	0	0	0	0	0	0	0
22	Y	0	0	0	0	0	0	0	0	0	0	0	0
23	Y	0	0	0	0	0	0	0	0.221	0	0	0	0

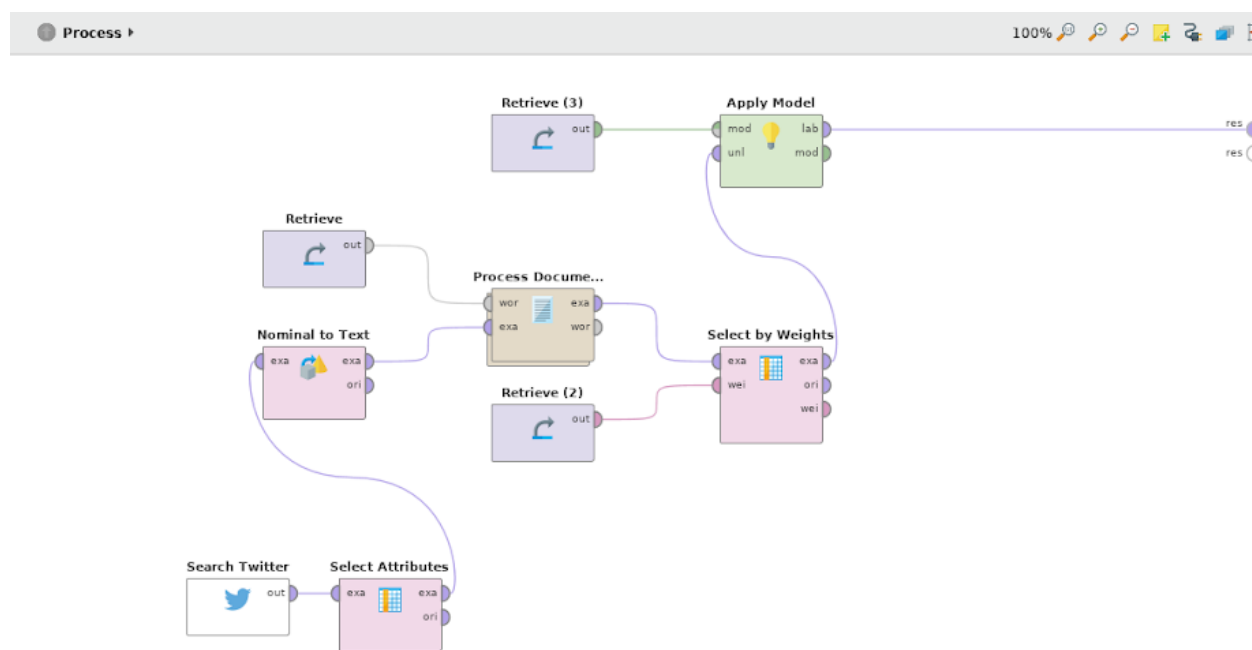
ExampleSet (200 examples, 1 special attribute, 440 regular attributes)

3.4.22 ประสิทธิภาพที่ได้จากการทำ Classification Model

accuracy: 70.50% +/- 1.58% (micro average: 70.50%)

	true Y	true N	class precision
pred. Y	141	59	70.50%
pred. N	0	0	0.00%
class recall	100.00%	0.00%	

3.4.23 ทำ Classify Tweet



3.4.24 ข้อมูลของ Tweet ที่ได้ดิงมา

Open in Turbo Prep Auto Model

Filter (100 / 100 examples): all

Row No.	Id	Created-At	From-User	From-Use...	To-User	To-User-Id	Language	Source	Text	Geo-Loca...	Geo-Loca...	Retweet...
1	112143595...	Apr 25, 20...	Bill McKibben	21786618	?	-1	en	<a href="h...	The fossil fu...	?	?	616
2	112114442...	Apr 25, 20...	Julia Hartley...	459390022	?	-1	en	<a href="h...	There's a w...	?	?	303
3	112142699...	Apr 25, 20...	Zaid Hamid	126966274	?	-1	en	<a href="h...	From Pakist...	?	?	233
4	112167694...	Apr 26, 20...	Max Tsui	763949599...	?	-1	en	<a href="h...	So we've en...	?	?	0
5	112167674...	Apr 26, 20...	Linda Mintun	804409343...	?	-1	en	<a href="h...	RT @billmc...	?	?	616
6	112167656...	Apr 26, 20...	Sharada Sri...	109707466...	?	-1	en	<a href="h...	RT @ErikSol...	?	?	38
7	112167651...	Apr 26, 20...	Jenni Clarke	780634555	?	-1	en	<a href="h...	I don't kno...	?	?	0
8	112167642...	Apr 26, 20...	xCuculainn...	1577059465	editfmah	288231300	en	<a href="h...	@editfmah ...	?	?	0
9	112167639...	Apr 26, 20...	Lynn	220996941	?	-1	en	<a href="h...	We prayed ...	?	?	0
10	112167639...	Apr 26, 20...	Ultimate Da...	947474480...	?	-1	en	<a href="h...	This issue o...	?	?	0
11	112167636...	Apr 26, 20...	Ethemcan	310295474	?	-1	en	<a href="h...	RT @comm...	?	?	29
12	112167631...	Apr 26, 20...	Margu501M...	2802155333	?	-1	en	<a href="h...	RT @Marko...	?	?	3
13	112167628...	Apr 26, 20...	xRobert M...	927850291...	?	-1	en	<a href="h...	RT @surfnu...	?	?	86
14	112167626...	Apr 26, 20...	□	102375607...	?	-1	en	<a href="h...	RT @ErikSol...	?	?	38
15	112167624...	Apr 26, 20...	📧James...	3117674405	EmmaSeng...	2596501422	en	<a href="h...	@EmmaSe...	?	?	0
16	112167620...	Apr 26, 20...	wufrose	847402223...	?	-1	en	<a href="h...	RT @sunflo...	?	?	91
17	112167620...	Apr 26, 20...	Columbia S...	22272603	?	-1	en	<a href="h...	RT @Dictio...	?	?	38
18	112167619...	Apr 26, 20...	Marie AuBu...	73797015	?	-1	en	<a href="h...	RT @Alex_V...	?	?	22
19	112167616...	Apr 26, 20...	Prof Nick Co...	1201902595	EarthB4worth	159480648	en	<a href="h...	@EarthB4w...	?	?	0
20	112167615...	Apr 26, 20...	Malik inzim...	101999515...	?	-1	en	<a href="h...	RT @TheTer...	?	?	131
21	112167609...	Apr 26, 20...	Alethea	111381481...	?	-1	en	<a href="h...	RT @Marko...	?	?	3
22	112167606...	Apr 26, 20...	World Trad...	108539418	?	-1	en	<a href="h...	RT @eagle...	?	?	10
23	112167605...	Apr 26, 20...	Julian Harvey	806666760...	?	-1	en	<a href="h...	RT @JohnSl...	?	?	94
24	112167605...	Apr 26, 20...	A Devoted ...	612931307	?	-1	en	<a href="h...	RT @SaferS...	?	?	3

ExampleSet (100 examples, 1 special attribute, 11 regular attributes)

3.4.25 ผลลัพธ์ของการทำ Classify Tweet

Open in Turbo Prep Auto Model

Filter (100 / 100 examples): all

Row No.	Id	predictio...	confiden...	confiden...	account	acid	action	adapt	addthi	adrian	aid	air	al
1	112143595...	Y	0.999	0.001	0	0	0	0	0	0	0	0	0
2	112114442...	Y	0.997	0.003	0	0	0	0	0	0	0	0	0
3	112142699...	Y	0.998	0.002	0	0	0	0	0	0	0	0	0
4	112167912...	Y	0.984	0.016	0	0	0	0	0	0	0	0	0
5	112167901...	Y	0.999	0.001	0	0	0	0	0	0	0	0	0
6	112167899...	Y	0.997	0.003	0	0	0	0	0	0	0	0	0
7	112167889...	Y	0.996	0.004	0	0	0	0	0	0	0	0	0
8	112167885...	Y	0.999	0.001	0	0	0	0	0	0	0	0	0
9	112167878...	Y	0.978	0.022	0	0	0	0	0	0	0	0	0
10	112167875...	Y	0.999	0.001	0	0	0	0	0	0	0	0	0
11	112167874...	Y	0.994	0.006	0	0	0	0	0	0	0	0	0
12	112167871...	Y	0.954	0.046	0	0	0	0	0	0	0	0	0
13	112167869...	Y	0.978	0.022	0	0	0	0	0	0	0	0	0
14	112167855...	Y	0.997	0.003	0	0	0	0	0	0	0	0	0
15	112167854...	Y	0.998	0.002	0	0	0	0	0	0	0	0	0
16	112167851...	Y	0.999	0.001	0	0	0	0	0	0	0	0	0
17	112167850...	Y	1.000	0.000	0	0	0	0	0	0	0	0	0
18	112167845...	Y	0.978	0.022	0	0	0	0	0	0	0	0	0
19	112167838...	Y	0.999	0.001	0	0	0	0	0	0	0	0	0
20	112167831...	Y	0.978	0.022	0	0	0	0	0	0	0	0	0
21	112167828...	Y	0.978	0.022	0	0	0	0	0	0	0	0	0
22	112167819...	Y	0.992	0.008	0	0	0	0	0	0	0	0	0
23	112167819...	Y	1.000	0.000	0	0	0	0	0	0	0	0	0
24	112167818...	Y	0.767	0.233	0	0	0	0	0	0	0	0	0

ExampleSet (100 examples, 4 special attributes, 440 regular attributes)

บทที่ 4

ผลการดำเนินโครงการ

4.1 ผลการดำเนินงาน

- ระบบ Big data Platform ที่ประกอบไปด้วย 1 Namenode และ 2 Datanode มีประสิทธิภาพในการวิเคราะห์ Big data ที่ดีกว่าคอมพิวเตอร์ทั่วไป
- ได้ระบบต้นแบบ Big Data Analytics Platform ที่ช่วยให้การวิเคราะห์ข้อมูลง่ายขึ้น
- ระบบที่ได้สามารถที่จะเพิ่มประสิทธิภาพได้ด้วยการเพิ่ม Datanode เพื่อกระจายการทำงาน
- ความแม่นยำของการทำ Data Science โดยใช้ RapidMiner อยู่ที่ 70%

accuracy: 70.50% +/- 1.58% (micro average: 70.50%)

	true Y	true N	class precision
pred. Y	141	59	70.50%
pred. N	0	0	0.00%
class recall	100.00% 0	0.00%	

บทที่ 5

สรุปผล อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการดำเนินโครงการ

- ได้ระบบต้นแบบ Big data analytics platform
- ระบบสามารถอ่านเขียนข้อมูลได้
- สามารถทำการวิเคราะห์ข้อมูลและทำ Data Science ได้
- หากประสิทธิภาพของ Namenode สูงกว่าจะทำงานได้ดีกว่านี้

5.2 ปัญหาและอุปสรรคในการดำเนินโครงการ

- 1.บทความที่เกี่ยวข้องโดยตรงในการติดตั้ง Cloudera บน CentOS มีน้อย ส่วนมากจะเป็นการติดตั้งบน Virtual Box
- 2.ในครั้งแรกและครั้งที่สองมีปัญหาในการเชื่อมต่อ RapidMiner เนื่องจาก Version ของ Spark ใน CDH เก่าเกินไป และใหม่เกินไป
- 3.เนื่องจากมีการใช้ RapidMiner ร่วมด้วย ทำให้การทำงานช้ากว่าที่คาด แต่การประมวลผลบน Hadoop เป็นไปอย่างปกติ

5.3 ข้อเสนอแนะ

จากปัญหาข้างต้น การติดตั้งระบบนั้นเครื่อง Namenode ควรมีหน่วยความจำเข้าถึงโดยสุ่ม (RAM) ที่มากกว่าเดิม เพื่อให้การใช้งานและการประมวลผลเป็นไปอย่างราบรื่น

แหล่งอ้างอิง

- [1] Big data analytics สำคัญยังไงและช่วยอะไรเราได้บ้าง. สืบค้นเมื่อ 22 พฤศจิกายน 2561 จาก <http://bigdataexperience.org/what-is-big-data-analytics/>
- [2] What is Hadoop. สืบค้นเมื่อ 22 พฤศจิกายน 2561 จาก <https://saixiii.com/what-is-hadoop/>
- [3] Micro Computer. คลาวเดอร์ (Cloudera). สืบค้นเมื่อ 22 พฤศจิกายน 2561 จาก <http://www.microvsmart.com/คลาวเดอร์-Cloudera/>
- [4] การประมวลผล Big Data ควรใช้เทคโนโลยีไหนดี. สืบค้นเมื่อ 22 พฤศจิกายน 2561 จาก <https://thanachart.org/2016/02/11/การประมวลผล-big-data-ควรใช้เทคโนโลยีไหนดี/>
- [5] การใช้งานและการติดตั้ง Apache Hive สำหรับ Big Data Processing. สืบค้นเมื่อ 22 พฤศจิกายน 2561 จาก <https://medium.com/@p.auoychai/การใช้งานและการติดตั้ง-apache-hive-สำหรับ-big-data-processing>
- [6] Data Mining Trend. บทที่ 1 แนะนำการใช้งาน RapidMiner Studio. สืบค้นเมื่อ 22 พฤศจิกายน 2561 จาก <http://dataminingtrend.com/2014/rapidminer-studio/chapter1/>
- [7] Manoj Kumar Danthala, “Tweet Analysis: Twitter Data processing Using Apache Hadoop.” In International Journal Of Core Engineering & Management (IJCEM), Volume 1, 2015
- [8] PrathyushaRani Merla, Yiheng Liang. “Data Analysis using Hadoop MapReduce Environment.” In Big Data (BIGDATA), 2017 International Conference on, pp 4783-4785.IEEE, 2017
- [9] Sudipto Shankar Dasgupta, Swaminathan Natarajan, Kiran Kumar Kaipa, Sujay Kumar Bhattacharjee, Arun Viswanathan. “Sentiment Analysis of Facebook Data using Hadoop based Open Source Technologies.”, IEEE, 2015
- [10] Apache Zookeeper. Welcome to Apache ZooKeeper™. สืบค้นเมื่อ 4 มกราคม 2562 จาก <https://zookeeper.apache.org/>
- [11] BigHadoop. Apache Spark – a Fast Big Data Analytics Engine. สืบค้นเมื่อ 4 มกราคม 2562 จาก <https://bighadoop.wordpress.com/2014/04/03/apache-spark-a-fast-big-data-analytics-engine/>